SOME MODERN APPLICATIONS OF MATHEMATICS

Stephen Barnett

Some Modern Applications of Mathematics

Some Modern Applications of Mathematics

Stephen Barnett
Department of Applied Mathematical Studies
University of Leeds

ELLIS HORWOOD
London New York London Toronto Sydney Tokyo Singapore
Madrid Mexico City Munich

First published 1995 by
Ellis Horwood Limited
Campus 400, Maylands Avenue
Hemel Hempstead
Hertfordshire, HP2 7EZ
A division of
Simon & Schuster International Group

Typeset in 10/12pt Times by Mathematical Composition Setters Ltd, Salisbury, Wiltshire

Printed and bound in Great Britain by
Redwood Books, Trowbridge, Wiltshire

Library of Congress Cataloging-in-Publication Data

Available from the publisher

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN 0-13-834094-3 (pbk)

1 2 3 4 5 99 98 97 96 95

Contents

Preface

For too long 'applied mathematics' in schools and universities has meant 'mechanics'. Although this area still has an important role to play, many students are turning away from it and are showing a growing interest in newer areas of applications of mathematics, such as those covered by the various syllabuses in A-level 'Decision mathematics'. These topics appeal to students because the applications are to problems arising in commerce, information technology and the environment, and generally do not involve a knowledge of physical principles.

The objective of this book is to communicate the flavour of some of these areas of recent applications of mathematics. It is intended to be read by students and their teachers on A-level courses at school or college, or in the first year of undergraduate degree courses. I have taught in the classroom all the material which is covered in this book, and it is interesting that this has steadily 'moved down' the curriculum. For example, I gave postgraduate lectures on linear programming in the early 1960s, yet this subject is now in Level 10 of the National Curriculum, as well as forming part of the A-level 'Decision mathematics' syllabus. In a similar way, I introduced control theory as an option for final-year mathematics undergraduates 25 years ago, whereas introductory courses are now taught at much lower levels. Part of the explanation for this moving down is due to the ready availability of computing power, enabling the use of efficient algorithms to solve problems. Throughout I have tried to emphasize discrete models using difference equations and matrix representations, and to reduce the emphasis on calculus and differential equations. This is for two reasons: first, students generally find discrete mathematics easier to grasp compared with the rather difficult concepts of calculus; and secondly, the amount of calculus being studied in schools is likely to decrease further in future. I have deliberately not included any discussion of computer software packages such as DERIVE or MAPLE. Not every student has ready access to these, and in any case software is revised or replaced so frequently that textbook treatments can quickly become out of date. Furthermore, I believe that at a beginning stage you learn and absorb concepts and techniques more readily by actually tackling problems with no more to help you than a good mathematical pocket calculator.

The book begins with a description of how measuring time in small steps leads to realistic models which do not involve calculus, and a number of applications using so-called difference equations are discussed. Methods for solving these equations are explained, and Chapter 1 closes with an account of how matrix algebra can be used to deal with more complicated problems, especially population models of the natural world.

Chapter 2 takes up the completely different topic of error-correcting codes. Applications include supermarket barcodes which identify products, the International Standard Book Number (ISBN), compact discs whose high quality of sound reproduction depends crucially on coding, and data sent from spacecraft which would be unintelligible but for appropriate coding. This is a particularly interesting area of contemporary applied mathematics, since so much of society now depends upon the accurate handling and transmission of information.

Another feature of the modern world is the use of computers to control the behaviour of engineering and other systems – for example, the automatic gearbox of a motor vehicle, the automatic landing of an aircraft, or putting a satellite into the correct orbit. An introduction to some of the mathematics involved is given in Chapter 3. In fact, because many control models involve dynamics this is one part of the book where a knowledge of the basics of Newton's laws of motion is useful. Although these models involve simple differential equations as well as difference equations, the main mathematical tool is matrix algebra. Required properties of matrices are developed as needed.

Finally, various aspects of the ever-present problem of optimizing the use of limited resources are investigated in Chapter 4, including linear programming, transportation problems and networks. The mathematics is at relatively simple levels except in the very last section on optimal control, which continues from Chapter 3 and can be regarded as something of a bridge to more advanced work.

Each chapter contains many worked examples, together with exercises which you should attempt to solve as you go through the book. At the end of each chapter there are further problems which are a bit more challenging, and you should at least skim over them, as they often contain further illustrative applications. Answers to all the numerical parts of the exercises and problems are provided. Teachers will be glad to know that a manual of written-out solutions to the exercises and problems is available on request from the publishers, free of charge. Students should *not*, in their own best interests, use it to cheat!

My style throughout is deliberately informal, with virtually nothing involving 'pure mathematics style' proofs. I introduce mathematical techniques as and when required for applications, so as to reduce the need to refer elsewhere. Some mathematical colleagues may well be dismayed by the absence of formal proofs, and may claim that I am encouraging sloppy ways of mathematical learning. However, for me mathematics has always come alive through its applications, and if the book manages to get this over to readers then I will be satisfied; rigorous developments which do not excite interest (except amongst a dedicated few) are what gives mathematics a bad name!

I wish to thank Brian Bunday, Tim Cronin, Mike Gover and Colin Storey for their helpful comments on first drafts, and Carolyn Barry for her excellent typing of the manuscript.

Stephen Barnett
Leeds, January 1995

1

Time Marches On

1.1 INTRODUCTION AND EXAMPLES

In real life, time is not measured continuously but in packets of a fixed amount, whether these be tenths of seconds, seconds, minutes, hours, days, months, years or whatever. For example, if you are ill with a fever then your temperature may be taken perhaps every hour – you certainly don't lie in bed with a thermometer permanently sticking out of your mouth so that a nurse can measure your temperature 'continuously'. In a similar way, if you have some money in a building society or bank to which interest is being added, this is only done at regular intervals, perhaps annually or monthly – rarely is the interest added on a daily basis, so it's no use checking every day to see whether your money's growing! Similarly, economic statistics such as the rate of inflation or the unemployment total are usually released at monthly or even quarterly intervals.

In this section we give some illustrations or situations where time is measured *discretely*, that is in finite, distinct amounts of 'steps' – so indeed 'time marches on', to quote the name of an American cinema newsreel popular in the 1940s and 1950s, and occasionally repeated on TV for historical interest.

■ EXAMPLE 1.1

A bank savings account pays interest at an annual percentage rate (APR) of r, which is 'compounded' n times per year, where n is an integer. For example, suppose that the APR is 5 compounded semi-annually, so that $n=2$. We use $x(0)$ to denote the amount which you put in to open the account, since this is when we start measuring time. At the end of six months, that is *one* time period, it's convenient to write $x(1)$ for the amount in the account. This will consist of the initial deposit together with the interest it has gained over the half-year, at *half* the annual rate (2.5%), so we have

$$x(1) = x(0) + \frac{2.5}{100}\,x(0)$$

$$= 1.025x(0)$$

If your money is left in for a further half-year, then 'compound' interest means that the whole of the new sum gains interest, not just the original deposit. Since *two* time periods have elapsed, the amount in the account is therefore

$$x(2) = 1.025\,x(1)$$

and substituting for $x(1)$ you can see that the amount in the account after a year (two time periods) has elapsed is

$$x(2) = (1.025)^2 x(0)$$

Let's see what happens in general when the year is divided up into n equal parts. At the *end* of each period the interest is added to the account at a rate of $r/100n$ (since r is a percentage) on the balance at the *beginning* of the period. Let $x(k)$ be the amount in the account at the end of the kth period, where k is the *time variable* which can take the values 1, 2, 3, ... After k time intervals the amount in the account is the previous balance together with the interest it has earned, so that

$$x(k) = x(k-1) + \frac{r}{100n}\,x(k-1)$$

$$= \left(1 + \frac{r}{100n}\right) x(k-1) \qquad (1.1)$$

Writing $\alpha = 1 + r/100n$ gives

$$x(k) = \alpha x(k-1), \quad k = 1, 2, 3, \ldots \qquad (1.2)$$

Equation (1.2) is an example of a *difference equation*, so called because it gives an expression for the difference between $x(k)$ and $x(k-1)$. The name *recurrence equation* (or *relation*) is also used, because the values of $x(k)$ can be computed recursively (i.e. one after the other) by simply substituting for k in (1.2) as follows:

$k=1$: $x(1) = \alpha x(0)$

$k=2$: $x(2) = \alpha x(1) = \alpha^2 x(0)$

$k=3$: $x(3) = \alpha x(2) = \alpha^3 x(0)$

and so on. You should be able to spot the pattern: the *general solution* of (1.2) is

$$x(k) = \alpha^k x(0), \quad k = 1, 2, 3, \ldots \tag{1.3}$$

It's worth pointing out that an equivalent version of (1.2), which is often used, is

$$x(k+1) = \alpha x(k), \quad k = 0, 1, 2, 3, \ldots$$

In this case we start counting at $k = 0$. There is no change in the general solution as it appears in (1.3).

An alternative, widely used notation is to write x_k instead of $x(k)$ to stand for the value of the variable after k time periods have elapsed.

So far we have assumed that the account is opened with an initial deposit $x(0)$, which is then left alone to accrue interest. However, it's usual to make frequent deposits and withdrawals from a bank account. Suppose that $u(k)$ is the *net* amount deposited during the kth time period, and this does not earn interest until the next period; if there is a net withdrawal then $u(k)$ is negative. The equation (1.1) then becomes

$$x(k) = \left(1 + \frac{r}{100n}\right) x(k-1) + u(k), \quad k = 1, 2, 3, \ldots$$

and this is an example of a general difference equation having the form

$$x(k) = \alpha x(k-1) + \beta u(k), \quad k = 1, 2, 3, \ldots \tag{1.4}$$

where α and β can take any constant values. Again, an alternative form of (1.4) is

$$x(k+1) = \alpha x(k) + \beta u(k+1), \quad k = 0, 1, 2, \ldots \tag{1.5}$$

The difference equations (1.4) and (1.5) are called *linear* because the variables are not raised to any powers. This is a mathematical definition which really doesn't convey much information, and because of the importance of the idea of *linearity* it's worth spending some time discussing it. Suppose we have a 'black box' represented in Figure 1.1. This box is a mysterious object — we don't know what is going on inside it; all we can do is put in 'things' (called *inputs*) and observe what comes out as a result (the *outputs*). The crucial *principle of linearity* is as follows.

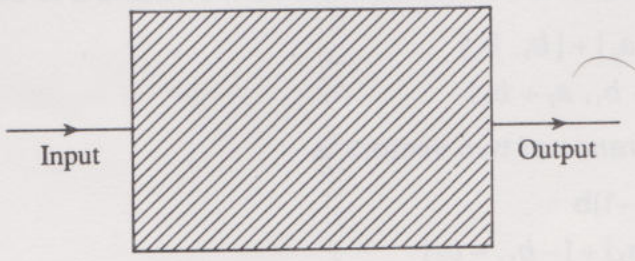


Figure 1.1

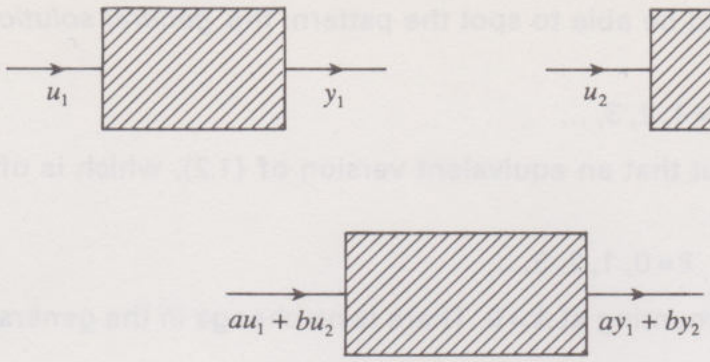


Figure 1.2

Suppose the response to an input u_1 is an output y_1, and for an input u_2 the output is y_2. Then the combination of inputs $au_1 + bu_2$ produces the *same* combination of outputs $ay_1 + by_2$, for any values of the constants a and b (Figure 1.2).

As an illustration, if, say, $a=2$ and $b=0$ then $2u_1$ produces $2y_1$ – that is, doubling the input results in a doubling of the output. This connects with the concept of a linear relationship between two variables which you will be familiar with as a straight line graph. However, the way we have described linearity above is much more useful, since we don't need to know what kind of mathematical processes are going on inside the 'box', so long as they obey the principle of linearity as described above.

■ *EXAMPLE 1.2*

Consider a certain linear 'system' which has inputs and outputs which are two-dimensional *vectors*. If you've not encountered the concept of vectors, a brief explanation is as follows. The notation $\mathbf{a} = [a_1, a_2]$ denotes a two-dimensional vector, which can be thought of as the line in the xy-plane from the origin to the point having coordinates $x = a_1$, $y = a_2$. For any scalar k the product $k\mathbf{a}$ is defined by

$$k[a_1, a_2] = [ka_1, ka_2]$$

which is a line from the origin to the point having coordinates $x = ka_1$, $y = ka_2$. If $\mathbf{b} = [b_1, b_2]$ is a second vector, then the *sum* of $\mathbf{a}$ and $\mathbf{b}$ is defined by

$$\begin{aligned}\mathbf{a} + \mathbf{b} &= [a_1, a_2] + [b_1, b_2] \\ &= [a_1 + b_1, a_2 + b_2]\end{aligned}$$

Hence the *difference* of two vectors is

$$\begin{aligned}\mathbf{a} - \mathbf{b} &= \mathbf{a} + (-1)\mathbf{b} \\ &= [a_1, a_2] + [-b_1, -b_2] \\ &= [a_1 - b_1, a_2 - b_2]\end{aligned}$$

Suppose it is found for our system that an input vector $\mathbf{u}_1 = [2, 1]$ produces an output vector $\mathbf{y}_1 = [3, 0]$, whereas $\mathbf{u}_2 = [1, 4]$ produces $\mathbf{y}_2 = [0, -1]$. Suppose that we then put in the combination of inputs

$$\begin{aligned} 2\mathbf{u}_1 - 3\mathbf{u}_2 &= 2[2, 1] - 3[1,4] \\ &= [4, 2] - [3, 12] \\ &= [1, -10] \end{aligned}$$

The corresponding output is, by the linearity principle, precisely the *same* combination of outputs, namely

$$\begin{aligned} 2\mathbf{y}_1 - 3\mathbf{y}_2 &= 2[3, 0] - 3[0, -1] \\ &= [6, 0] - [0, -3] \\ &= [6, 3] \end{aligned}$$

In Example 1.2 we referred to a 'system', this could be a mathematical description involving difference equations, or differential equations, or matrices — precisely what is involved is irrelevant.

EXERCISE 1.1 An old will has just been found showing that your great-grandfather, who died 60 years ago, left you £5 (a fair sum of money back then!) which has been earning interest at an APR of seven compounded quarterly (i.e. every 3 months). How much will you now get?

EXERCISE 1.2 A saving account pays 5% compounded semi-annually. The initial deposit is £1000, and net deposits during successive half-years are £476, £355, −£217, £727. Determine the balance in the account at the end of 2 years.

EXERCISE 1.3 A linear system has a vector response $[1, -19]$ when the vector input is $[1, 1]$, and a response $[1, -31]$ when the input is $[2, 1]$. Express an input $[1, 0]$ as a linear combination of the two inputs — that is, find constants a and b such that

$$[1, 0] = a[1, 1] + b[2, 1]$$

Hence determine the response of the system when the input is $[1, 0]$.

EXERCISE 1.4 If the annual rate of inflation in the economy is 5%, this means that at the end of the year you need £1.05 in order to buy what £1 would have purchased at the beginning of the year. If this rate of inflation continues for 10 years, how much do you need in order to have the equivalent of £100 purchasing power at the start of the next decade?

■ *EXAMPLE 1.3 Fish aquarium model*

In an aquarium it is important to prevent the build-up of a high concentration of salt which is dangerous to fish. Suppose we try to do this in the following way. We notice that 1 unit of water evaporates during the week. As a restorative

tactic, at the end of every week we remove a further 2 units of water, and then add 3 units of fresh water. Let n be the total number of units of water in the aquarium, let s be the concentration of salt per unit of fresh water, and let $x(k)$ denote the total amount of salt in the aquarium at the end of the kth week, after the water level has been brought back to normal. At the beginning ($k=0$) we therefore have $x(0) = ns$. After 1 week there are $n-1$ units of water left, and the salt content per unit is therefore $x(0)/(n-1)$. We remove two of these units and add three fresh ones, so the net amount of salt at the end of week 1 is

$$x(1) = x(0) - \frac{2x(0)}{n-1} + 3s$$

initial amount of salt; salt in removed units; salt in fresh units added

Similarly, at the end of week k the salt content is

$$x(k) = x(k-1) - \frac{2x(k-1)}{n-1} + 3s$$

salt at beginning of week k; salt in removed units; salt in fresh units added

Simplifying this equation gives

$$x(k) = \frac{n-3}{n-1}\,x(k-1) + 3s, \quad k = 1, 2, 3, \ldots \tag{1.6}$$

which you can see has exactly the form (1.4) with $\alpha = (n-3)/(n-1)$, $\beta u(k) = 3s$ (for all k).

■ *EXAMPLE 1.4 Rabbit population model*

This description of a rabbit population originates with an Italian mathematician called Fibonacci in the early part of the thirteenth century, and because of its long history has been widely studied. We make the following assumptions:

(i) begin with a pair of newborn rabbits (one male, one female);
(ii) a newborn pair matures (i.e. becomes adult) after 1 month, and produces the first offspring at 2 months of age;
(iii) a pair (one male, one female) is born to each pair of adult rabbits at the end of every month;
(iv) once paired, rabbits remain faithful to each other and do not die!

These assumptions are, of course, not actually attainable in practice — for example, (iv) assumes an unlimited food supply together with immortality, so we can think of a very large grassy island free of predators as being a rabbit's idea of heaven! Let $x(k)$ be the number of pairs of rabbits present at the end of the kth month, beginning with $x_0 = 1$ (we shall from now on use the neater notation x_k instead of $x(k)$). After 1 month, by assumption (ii) this pair has matured but has no offspring, so $x_1 = 1$. After a second month has passed, a

pair of offspring is produced, so there are in total two pairs (i.e. $x_2 = 2$). At the end of the third month the original pair has produced another pair of offspring, and their first-born has matured, so in total $x_3 = 3$. This is represented in Figure 1.3, where the next few months are also shown.

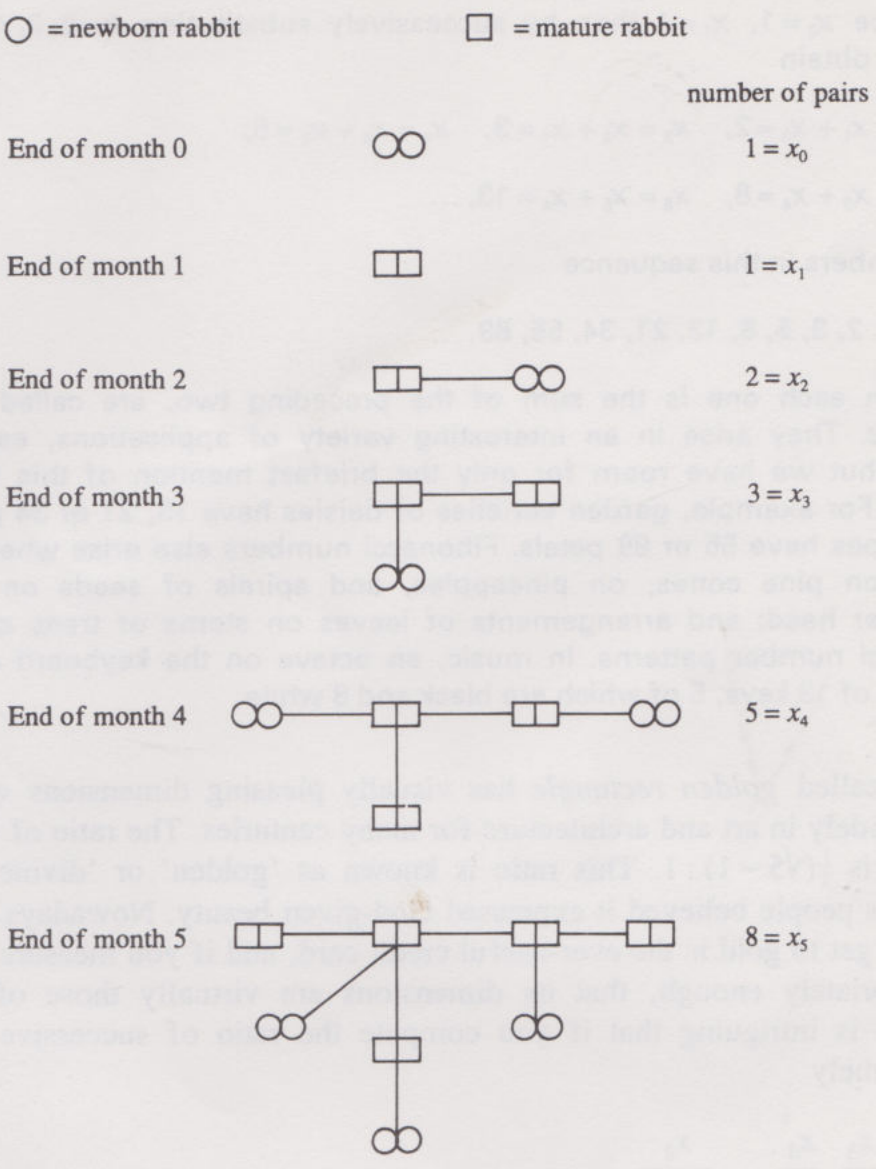


Figure 1.3

In general at the end of the kth month

$$x_k = x_{k-1} + x_{k-2} \tag{1.7}$$

total number of pairs at end of month k = number of *adult* pairs at end of month k + number of *newborn* pairs at end of month k

The way (1.7) is built up is because of assumptions (ii) and (iii): the *total* number of pairs at the end of month $k-1$ (i.e. x_{k-1}) becomes the number of

adult pairs 1 month later, and the x_{k-2} pairs at the end of month $k-2$ all produce offspring 2 months later. For example, in Figure 1.3 you can see that $x_5 = x_4 + x_3$: all the shapes in the row for the end of month 4 have become squares (adults) in the row below, and the three pairs shown at the end of month 3 all produce offspring (circles) at the end of month 5.

Since $x_0 = 1$, $x_1 = 1$ then by successively substituting $k = 2, 3, 4, 5, \ldots$ into (1.7) we obtain

$$x_2 = x_1 + x_0 = 2, \quad x_3 = x_2 + x_1 = 3, \quad x_4 = x_3 + x_2 = 5,$$

$$x_5 = x_3 + x_4 = 8, \quad x_6 = x_5 + x_4 = 13, \ldots$$

The numbers in this sequence

1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, ...

in which each one is the sum of the preceding two, are called *Fibonacci numbers*. They arise in an interesting variety of applications, especially in nature, but we have room for only the briefest mention of this fascinating subject. For example, garden varieties of daisies have 13, 21 or 34 petals, and other types have 55 or 89 petals. Fibonacci numbers also arise when counting spirals on pine cones, on pineapples, and spirals of seeds on a mature sunflower head; and arrangements of leaves on stems of trees also exhibit Fibonacci number patterns. In music, an octave on the keyboard of a piano consists of 13 keys, 5 of which are black and 8 white.

The so-called *golden rectangle* has visually pleasing dimensions which have been used widely in art and architecture for many centuries. The ratio of the lengths of its sides is $\frac{1}{2}(\sqrt{5}-1) : 1$. This ratio is known as 'golden' or 'divine', since in ancient times people believed it expressed God-given beauty. Nowadays the closest most people get to gold is the ever-useful credit card, and if you measure one you'll find, appropriately enough, that its dimensions are virtually those of a golden rectangle! It is intriguing that if you compute the ratio of successive Fibonacci numbers, namely

$$\frac{x_1}{x_2}, \frac{x_2}{x_3}, \frac{x_3}{x_4}, \frac{x_4}{x_5}, \ldots, \frac{x_k}{x_{k+1}}, \ldots$$

then as k gets larger and larger this ratio gets closer and closer to the number

$$\tfrac{1}{2}(\sqrt{5}-1) = 0.618\,033\,98\ldots$$

We'll see in Section 4.1, Chapter 4, that Fibonacci numbers are also useful in optimization (finding maximum or minimum values of functions).

EXERCISE 1.5 Using a calculator, work out the values of the ratio of Fibonacci numbers x_k/x_{k+1} for $k = 1, 2, 3, \ldots, 12$.

EXERCISE 1.6 Verify by direct substitution that an expression for the kth Fibonacci number in the form

$$x_k = \frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^{k+1} - \left(\frac{1-\sqrt{5}}{2}\right)^{k+1}\right], \quad k = 0, 1, 2, 3, \ldots \tag{1.8}$$

satisfies equation (1.7). Verify that it gives the correct values for $k = 0, 1, 2, 3$.

It is interesting that despite the presence of $\sqrt{5}$ in (1.8), the formula always gives an integer value for x_k (see Problem 1.7). We shall see later in Section 1.2 (Example 1.8) how to derive (1.8).

EXERCISE 1.7 Notice that for large values of k, since $(1-\sqrt{5})/2 \simeq -0.62$, the second term inside the square brackets in (1.8) becomes extremely small, and therefore

$$x_k \simeq \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^{k+1}$$

Hence deduce that for sufficiently large values of k

$$\frac{x_k}{x_{k+1}} \simeq \frac{2}{1+\sqrt{5}} = \frac{\sqrt{5}-1}{2}$$

EXERCISE 1.8 Denote the Fibonacci numbers by $f_0 = 1$, $f_1 = 1$, $f_2 = 2$, $f_3 = 3$ and so on.

(a) Verify that

$$f_0 + f_1 = f_3 - 1, \quad f_0 + f_1 + f_2 = f_4 - 1$$

and prove by induction that in general

$$f_0 + f_1 + f_2 + \cdots + f_n = f_{n+2} - 1$$

(See the appendix to this chapter for a description of the method of proof by induction, if you are not already familiar with it.)

(b) Similarly, verify that

$$f_0 + f_2 = f_3, \quad f_0 + f_2 + f_4 = f_5$$

and prove by induction that in general

$$f_0 + f_2 + f_4 + \cdots + f_{2n} = f_{2n+1}$$

EXERCISE 1.9 Consider the following model of a bee population in a hive. Unfertilized eggs laid by a queen bee hatch into males, and fertilized eggs hatch into females. In other words, male bees do not have fathers. The queen bee is able to regulate the gender of her offspring to meet the needs of the hive according to information supplied by her attendants. Thus the male's only function in the hive is his role in the production of females!

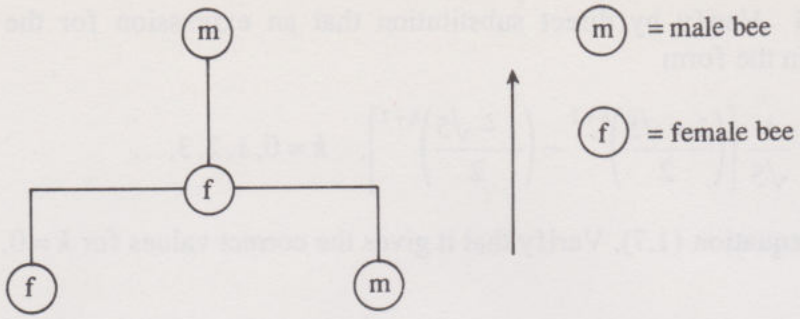


Figure 1.4

The ancestry of a single male can be traced using an appropriate diagram, which begins as shown in Figure 1.4. This is to be read upwards (in the direction of the arrow) and shows that the male in question had a mother only, together with a grandmother and grandfather. The previous stage is shown in Figure 1.5. Notice that the female parentage always branches into two.

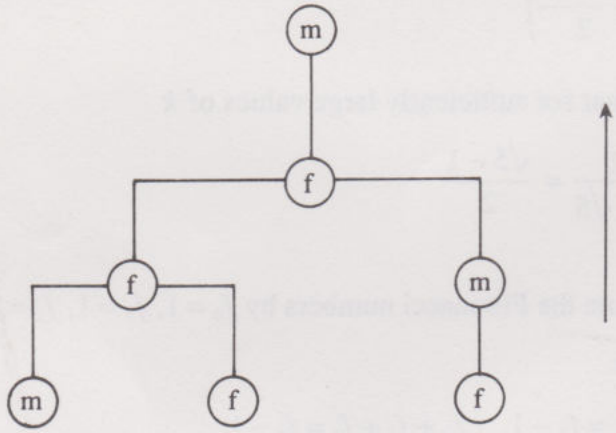


Figure 1.5

Extend the diagram in Figure 1.5 by going back two more generations.

Now count the numbers of bees at each level in the ancestry diagram. For example, in Figure 1.5, this time starting at the top and going downwards, we have the following:

Male	Female	Total
1	0	1
0	1	1
1	1	2
1	2	3

Extend the table to cover your diagram, and confirm that the numbers in the 'total' column are the Fibonacci numbers. Prove that this is true in general. (Hint: if x_k is the number in row k in the 'total' column, show that the numbers in the 'male' and 'female' columns in the same row are respectively x_{k-2} and x_{k-1}.)

The Fibonacci equation (1.7) is different from the equations in our earlier examples, in that it contains the variables at *three* moments in time, namely x_{k+2},

x_{k+1} and x_k. Since this involves differences at *two* units of time (i.e. $x_{k+2} - x_k$) it is called a *second-order equation*, in contrast with the simpler type in (1.4) which is called *first-order*. A useful way of expressing second-order equations is to employ the notation and ideas of matrices. Let's consider a more general form than (1.7), which we can write as

$$x_{k+2} = ax_{k+1} + bx_k, \quad k = 0, 1, 2, \ldots \tag{1.9}$$

where a and b are constants. In order to start things off with second-order equations we need to be given *two* known values, usually those of x_0 and x_1; for example, in the Fibonacci equation we had $x_0 = 1$, $x_1 = 1$. Define a 'new' variable by $y_0 = x_1$, $y_1 = x_2$, $y_2 = x_3$, ... so that in general

$$y_k = x_{k+1}, \quad k = 0, 1, 2, 3, \ldots \tag{1.10}$$

and also $y_{k+1} = x_{k+2}$. Substituting for this new variable into (1.9) gives

$$y_{k+1} = bx_k + ay_k \tag{1.11}$$

We can combine together (1.10) and (1.11) into a matrix form

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ b & a \end{bmatrix} \begin{bmatrix} x_k \\ y_k \end{bmatrix} \tag{1.12}$$

Provided that you are acquainted with the basics of matrix algebra (if not, don't worry, as we'll give an explanation in Section 1.4) then you'll realize that on expanding out the product in (1.12) we simply get back to (1.10) and (1.11). Equation (1.12) is itself a special case of an equation which is yet more general:

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = A \begin{bmatrix} x_k \\ y_k \end{bmatrix}, \quad k = 0, 1, 2, \ldots \tag{1.13}$$

where A can be an *arbitrary* 2×2 matrix. It is interesting to consider a situation where (1.13) arises *directly*, rather than as in (1.12) by converting the single second-order equation (1.9) into matrix form.

■ *EXAMPLE 1.5 Bird population model*

Let's consider only the *females* in a population consisting of a single bird species. This is assumed to obey the following rules, which have been constructed on the basis of observations made over a number of years:

(i) a proportion α of juvenile females born in one year survives to become adults in the following spring;
(ii) each surviving adult female lays eggs in spring to produce an average of γ juveniles by the next spring;
(iii) adults die during the year for various reasons, a proportion β surviving from one spring to the next.

Let x_k, y_k denote the numbers of juvenile and adult females respectively in year k. Then assumption (ii) states that the number of juveniles in year $k+1$ is

$$x_{k+1} = \gamma y_k, \quad k = 0, 1, 2, 3, \ldots$$

The other two assumptions tell us that the number of adult females in year $k+1$ is

$$y_{k+1} = \underbrace{\alpha x_k}_{\substack{\text{number of juveniles from year } k \text{ who achieved adulthood}}} + \underbrace{\beta y_k}_{\substack{\text{number of adults surviving from year } k}}, \quad k = 0, 1, 2, \ldots$$

Combining these two equations together gives us the matrix form (1.13):

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & \gamma \\ \alpha & \beta \end{bmatrix}}_{A} \begin{bmatrix} x_k \\ y_k \end{bmatrix}$$

The assumptions (i)–(iii) are of course idealized. In real life the birth and death rates vary with the size of population due to limited food supplies and overcrowding, so that α, β and γ are not constants.

Economics is a fertile field for discrete time models since the relevant data (e.g. profits, investments, income, etc.) are obtained at well-defined instants of time – usually weekly, quarterly or yearly.

EXERCISE 1.10 A very simple model of a national economy assumes that in year k the national income I_k is equal to $C_k + P_k + G_k$, where C_k is consumer expenditure (e.g. on consumer goods), P_k is private investment (e.g. on manufacturing equipment) and G_k is government expenditure (e.g. on education and health). The following assumptions are based on investigation of past data:

(i) consumer spending is proportional to national income in the *previous* year, that is

$$C_k = \alpha I_{k-1}$$

where α is a constant;

(ii) private investment is proportional to the change in consumer spending over the *previous* year, i.e.

$$P_k = \beta(C_k - C_{k-1})$$

where β is a constant.

Show that I_k satisfies the second-order difference equation

$$I_{k+2} - \alpha(1+\beta)I_{k+1} + \alpha\beta I_k = G_{k+2}, \quad k = 0, 1, 2, \ldots$$

EXERCISE 1.11 A model for population movements into and out of California is based on evidence that 10% of the United States population outside California moves into that state every year, whereas 20% of the population of California moves out every year.

Let x_k, y_k be the numbers of people living respectively outside and inside California in year k. Derive a model in the form (1.13), and state the matrix A.

By using the substitution

$$x_k = 2u_k + v_k, \quad y_k = u_k - v_k$$

show that

$$2u_{k+1} + v_{k+1} = 2u_k + 0.7v_k$$
$$u_{k+1} - v_{k+1} = u_k - 0.7v_k$$

Hence show that

$$u_k = u_0, \quad v_k = (0.7)^k v_0, \quad \text{for } k = 1, 2, 3, \ldots$$

Finally, deduce that after sufficient years have passed so that $(0.7)^k \simeq 0$, then under the stated assumptions one-third of the population of the United States would be living in California.

1.2 SOLUTION OF DIFFERENCE EQUATIONS

We now study in more detail how to solve linear difference equations. We saw in the previous section that the general solution of the first-order equation

$$x_{k+1} = \alpha x_k, \quad k = 0, 1, 2, 3, \ldots \tag{1.14}$$

is $x_k = \alpha^k x_0$. It is often important in applications to know what happens to x_k as k becomes larger and larger. We use the notation $k \to \infty$ to mean that k increases indefinitely. Clearly if α is a real number whose magnitude is less than one then α^k gets smaller and smaller as k increases. We write

$$\alpha^k \to 0 \quad \text{as} \quad k \to \infty$$

(read as: 'α^k tends to zero as k tends to infinity') to mean that the magnitude of α^k can be made smaller than any positive quantity you care to name, simply by making k large enough. To take a simple example, if $\alpha = 0.1$ then we can make α^k smaller than 10^{-20}, say, by taking $k > 10^{20}$. Conversely, if α has magnitude greater than one, then α^k gets bigger and bigger as k increases, and we now write

$$\alpha^k \to \infty \quad \text{as} \quad k \to \infty$$

('α^k tends to infinity as k tends to infinity') to mean that the magnitude of α^k can be made larger than any specified positive quantity simply by taking k large enough.

We use the notation $|\alpha|$, the *modulus* of α, to denote the magnitude, or numerical value, of α irrespective of its sign, so $|\alpha| < 1$ means that $-1 < \alpha < 1$. If α is a complex number with real and imaginary parts u and v, so that $\alpha = u + \mathrm{i}v$ where $\mathrm{i}^2 = -1$, then α can be represented in the so-called *Argand diagram* shown in Figure 1.6. The *modulus* $|\alpha|$ of α is the distance r from the origin to the point with coordinates u and v, so

$$|\alpha| = r = (u^2 + v^2)^{1/2} \tag{1.15}$$

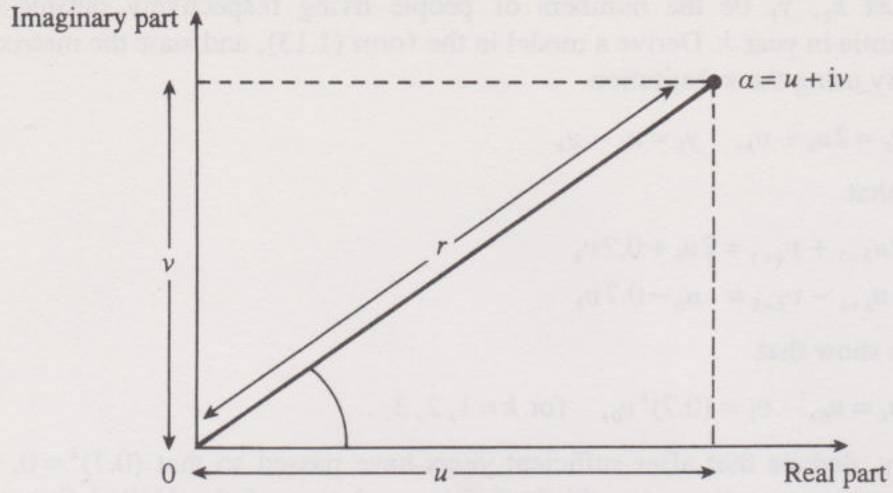


Figure 1.6

The *argument* θ is the angle shown in Figure 1.6, and $u = r\cos\theta$, $v = r\sin\theta$ so that

$$\alpha = r(\cos\theta + i\sin\theta)$$

A famous result about complex numbers states that for any angle θ (in radians)

$$e^{i\theta} = \cos\theta + i\sin\theta$$

so we can write $\alpha = re^{i\theta}$. Moreover, for any positive integer k it follows that

$$\begin{aligned}(e^{i\theta})^k &= e^{ik\theta} \\ &= \cos k\theta + i\sin k\theta\end{aligned}$$

Let's look at what happens to the form of the solution of (1.14) when α is complex:

$$\begin{aligned}x_k &= \alpha^k x_0 \\ &= (re^{i\theta})^k x_0 \\ &= r^k(\cos k\theta + i\sin k\theta)x_0\end{aligned}$$

Now, for *any* angle θ we always have

$$-1 \leqslant \cos k\theta \leqslant 1, \quad -1 \leqslant \sin k\theta \leqslant 1$$

It therefore follows that as $k \to \infty$ then $x_k \to 0$ whenever $r < 1$, and $x_k \to \infty$ if $r > 1$.

We can combine together the results for when α is real or complex, and say that as $k \to \infty$ then $x_k \to 0$ when $|\alpha| < 1$, and $x_k \to \infty$ if $|\alpha| > 1$, with the understanding that when α is a complex number its modulus is defined by (1.15).

EXERCISE 1.12 Investigate the solution of (1.14) when $\alpha = \pm 1$, or $\alpha = \pm i$.

We now move on to more general first-order equations, first seen in (1.4) and (1.5), where there is an extra term on the right-hand side of (1.14). Let's look first at

$$x_{k+1} = \alpha x_k + c, \quad k = 0, 1, 2, \ldots \tag{1.16}$$

where c is a constant. Substitute consecutive values of k into (1.16), starting with $k=0$:

$$x_1 = \alpha x_0 + c$$

$$x_2 = \alpha x_1 + c = \alpha(\alpha x_0 + c) + c = \alpha^2 x_0 + \alpha c + c$$

$$x_3 = \alpha x_2 + c = \alpha^3 x_0 + \alpha^2 c + \alpha c + c$$

$$x_4 = \alpha x_3 + c = \alpha^4 x_0 + (\alpha^3 + \alpha^2 + \alpha + 1)c$$

You should be able to spot the pattern: for a general value of k this is

$$x_k = \alpha^k x_0 + (\alpha^{k-1} + \alpha^{k-2} + \cdots + \alpha^2 + \alpha + 1)c \tag{1.17}$$

The term within brackets in this expression is

$$S_k = 1 + \alpha + \alpha^2 + \cdots + \alpha^{k-1} \tag{1.18}$$

and is called a *geometric series*, each term in it being α times the previous one, starting with 1 (there are k terms altogether). Multiplying (1.18) by α gives

$$S_k \alpha = \alpha + \alpha^2 + \cdots + \alpha^k \tag{1.19}$$

and subtracting (1.19) from (1.18) produces

$$S_k - S_k \alpha = 1 - \alpha^k \tag{1.20}$$

which can be simplified to

$$S_k = \frac{1-\alpha^k}{1-\alpha}$$

This is a well-known formula for the sum of a geometric series, but you have probably noticed that it doesn't work if $\alpha = 1$, since we can't then divide by $1-\alpha$; indeed (1.20) simply says that $0=0$! However, when $\alpha = 1$ each term in the series in (1.18) is itself 1, so that $S_k = k$. Putting these facts together, the general solution (1.17) of (1.16) is

$$x_k = \alpha^k x_0 + \begin{cases} \dfrac{(1-\alpha^k)c}{1-\alpha}, & \alpha \neq 1 \\ kc, & a = 1 \end{cases} \tag{1.21}$$

We have seen that if $|\alpha| < 1$ then $\alpha^k \to 0$ as $k \to \infty$. Hence in (1.21) when $|\alpha| < 1$ the terms involving α^k become insignificant for large enough k, so that x_k approaches $c/(1-\alpha)$.

■ EXAMPLE 1.6

Let's return to the aquarium model described in Example 1.3. Comparing equations (1.6) and (1.16) we see that

$$\alpha = \frac{n-3}{n-1}, \quad c = 3s$$

so that $0 < \alpha < 1$. Therefore as k becomes large, the total amount of salt x_k in the aquarium approaches

$$\frac{c}{1-\alpha} = \frac{3s(n-1)}{2}$$

Since the aquarium contains n units of water, the concentration of salt is $3s(n-1)/2n$. If the tank is large then $(n-1)/2n \simeq \frac{1}{2}$, so the concentration of salt becomes approximately $3s/2$ after a long period of time has elapsed – that is, 50% higher than the original concentration.

EXERCISE 1.13 Use (1.21) to determine the solution of the following equations, subject to the given condition:

(a) $x_{k+1} - 2x_k = 4, \; x_0 = 3$
(b) $x_{k+1} - x_k = -2, \; x_0 = 7$

In each case, verify by direct substitution into the equation that your solution is correct.

EXERCISE 1.14 Suppose that a different scheme is operated for the aquarium model in Example 1.3, in which at the end of each week we remove a single unit of water, and then add 2 units of fresh water so as to bring the level back to normal (recall that 1 unit of water evaporates during the week). Obtain the difference equation in this case, corresponding to (1.6). Show that in this case, again assuming that n is large, the concentration of salt will effectively double over a long time period.

Let's now increase the level of difficulty a further notch by taking the extra term in the equation to be k, instead of a constant:

$$x_{k+1} = \alpha x_k + k, \quad k = 0, 1, 2, \ldots \tag{1.22}$$

To get the solution, again substitute consecutive values of k:

$$x_1 = \alpha x_0$$
$$x_2 = \alpha x_1 + 1 = \alpha(\alpha x_0) + 1 = \alpha^2 x_0 + 1$$
$$x_3 = \alpha x_2 + 2 = \alpha^3 x_0 + \alpha + 2$$
$$x_4 = \alpha x_3 + 3 = \alpha^4 x_0 + \alpha^2 + 2\alpha + 3$$

Here you should be able to see that the pattern for a general value of k is

$$x_k = \alpha^k x_0 + \alpha^{k-2} + 2\alpha^{k-3} + 3\alpha^{k-4} + \cdots + (k-2)\alpha + k - 1 \tag{1.23}$$

The expression (1.23) is the general solution of (1.22), as can be verified by direct substitution. When $\alpha = 1$ then (1.23) reduces to

$$\begin{aligned} x_k &= x_0 + 1 + 2 + 3 + \cdots + (k-1) \\ &= x_0 + \tfrac{1}{2}k(k-1) \end{aligned}$$

where we have used a standard formula for the sum of the integers from 1 to $k-1$ (see (A1) in the appendix to this chapter).

EXERCISE 1.15 A standard identity (see Problem 1.10) states that

$$\theta + 2\theta^2 + 3\theta^3 + \cdots + (k-1)\theta^{k-1} = \frac{\theta[1 - k\theta^{k-1} + (k-1)\theta^k]}{(1-\theta)^2}, \quad \theta \neq 1$$

Use this to show that (1.23) can be rewritten as

$$x_k = \alpha^k x_0 + \frac{\alpha^k - k\alpha + k - 1}{(1-\alpha)^2}, \quad \alpha \neq 1 \tag{1.24}$$

EXERCISE 1.16 Determine the solution of each of the following equations valid for $k = 0, 1, 2, 3, \ldots$, and subject to the stated condition:

(a) $x_{k+1} = 3x_k + k$, $x_0 = 4$

(b) $x_{k+1} + 2x_k = 2 - k$, $x_0 = 2$

(c) $x_{k+1} - x_k = k + 2$, $x_0 = -1$

In each case, verify that your answer is correct by checking that it does indeed satisfy the given equation. Notice that in (b) and (c) you need to use the principle of linearity: regard the right-hand side as a sum of two 'inputs', and the corresponding solution as the 'output'.

EXERCISE 1.17 Consider a roll of kitchen foil which is wound around a cardboard core cylinder of radius 3 cm. The foil is 0.005 cm thick, so when the foil is wrapped k times around the core the outer radius of the roll is $3 + 0.005k$ cm. Let x_k be the total length of foil when it is wrapped k times around the core, with $x_0 = 0$.

(a) Show that

$$x_{k+1} = x_k + 6\pi + 0.01\pi k, \quad k = 0, 1, 2, \ldots$$

and solve this equation to obtain an expression for x_k.

(b) When the outer radius of the roll is 3.2 cm, what is the total length of foil?

(c) What is the outer radius of the roll when it holds a total length of 209 cm?

We can now turn to solving second-order equations in the form (1.9), which we'll rearrange as

$$x_{k+2} - ax_{k+1} - bx_k = 0, \quad k = 0, 1, 2, \ldots \tag{1.25}$$

In view of the solution of the first-order case (1.14), which we found to be in the form $x_k = \alpha^k x_0$, it seems reasonable to see if this also works for (1.25). To avoid

confusion, we'll use a parameter λ instead of α, and try $x_k = \lambda^k$, where clearly $\lambda \neq 0$, since $x_k = 0$ is a trivial and uninteresting solution. Substituting this into (1.25) gives

$$\lambda^{k+2} - a\lambda^{k+1} - b\lambda^k = 0$$

which can be factorized as

$$\lambda^k(\lambda^2 - a\lambda - b) = 0$$

We have rejected $\lambda = 0$ as a possibility, so we must have

$$\lambda^2 - a\lambda - b = 0 \tag{1.26}$$

If this quadratic equation has two roots λ_1, λ_2 different from each other, then we've shown that each of

$$x_k = c_1\lambda_1^k, \quad x_k = c_2\lambda_2^k$$

is a solution of (1.25), for any constants c_1 and c_2. It therefore follows by the principle of linearity that

$$x_k = c_1\lambda_1^k + c_2\lambda_2^k \tag{1.27}$$

is also a solution, and this is in fact the most general form. It contains *two* constants c_1 and c_2 whose values are determined by using two given conditions, usually specified values of x_0 and x_1.

■ EXAMPLE 1.7

We solve the equation

$$x_{k+2} + 5x_{k+1} + 6x_k = 0 \tag{1.28}$$

subject to $x_0 = 2$, $x_1 = 3$. The quadratic equation (1.26) is

$$\lambda^2 + 5\lambda + 6 = 0$$

which factorizes to

$$(\lambda + 2)(\lambda + 3) = 0$$

so the roots are $\lambda_1 = -2$, $\lambda_2 = -3$. The solution of (1.28) is therefore

$$x_k = c_1(-2)^k + c_2(-3)^k \tag{1.29}$$

To find the values of c_1 and c_2 we substitute $k = 0$, $k = 1$ into (1.29) and use $(-2)^0 = 1$, $(-3)^0 = 1$ to get

$$2 = c_1 + c_2$$

$$3 = -2c_1 - 3c_2$$

The solution of these equations is $c_1 = 9$, $c_2 = -7$, so the required solution of (1.28) is therefore

$$x_k = 9(-2)^k - 7(-3)^k, \quad k = 0, 1, 2, \ldots$$

■ *EXAMPLE 1.8*

We can now derive the solution given in (1.8) of the Fibonacci equation (1.7), which we write here as

$$x_{k+2} - x_{k+1} - x_k = 0, \quad k = 0, 1, 2, \ldots$$

The equation (1.25) is

$$\lambda^2 - \lambda - 1 = 0$$

Its roots, using the standard formula for solving a quadratic equation, are

$$\lambda_1, \lambda_2 = \frac{1 \pm \sqrt{(1+4)}}{2} = \frac{1 \pm \sqrt{5}}{2}$$

so that

$$\lambda_1 = \tfrac{1}{2}(1+\sqrt{5}), \quad \lambda_2 = \tfrac{1}{2}(1-\sqrt{5})$$

and the solution is therefore

$$x_k = c_1\left(\frac{1+\sqrt{5}}{2}\right)^k + c_2\left(\frac{1-\sqrt{5}}{2}\right)^k \tag{1.30}$$

It now remains to find the values of c_1 and c_2 using $x_0 = 1$, $x_1 = 1$. This gives the simultaneous equations

$$1 = c_1 + c_2$$

$$1 = c_1\left(\frac{1+\sqrt{5}}{2}\right) + c_2\left(\frac{1-\sqrt{5}}{2}\right)$$

and you should check that their solution is

$$c_1 = (1+\sqrt{5})/2\sqrt{5}, \quad c_2 = (\sqrt{5}-1)/2\sqrt{5}$$

Substituting these values into (1.30) then gives the solution we quoted earlier in (1.8).

EXERCISE 1.18 Determine the solution of each of the following second-order equations, subject to the stated conditions:

(a) $x_{k+2} + 7x_{k+1} + 12x_k = 0$, $x_0 = 2$, $x_1 = 3$
(b) $x_{k+2} - 4x_{k+1} + 5x_k = 0$, $x_0 = 2$, $x_1 = 4 + 4\mathrm{i}$

with $k = 0, 1, 2, 3, \ldots$ in each case.

You might be wondering at this stage how we were able to exclude $\lambda = 0$ definitely as a possible value, when you know that a quadratic equation like (1.26) can sometimes have a zero root. However, if $\lambda = 0$ is substituted as a root into (1.26), then this requires that $b = 0$, in which case the difference equation (1.25) becomes

$$x_{k+2} - ax_{k+1} = 0$$

This is no longer a second-order equation, merely a first-order difference equation (admittedly in a slightly disguised form) and its solution is just $x_k = a^k x_0$, as before.

There is one situation, though, that we must examine in extra detail: this is when the quadratic equation (1.26) has two *equal* roots, say $\lambda = \lambda_3$, in which case the solution (1.27) becomes

$$\begin{aligned} x_k &= (c_1 + c_2)\lambda_3^k \\ &= c_3\lambda_3^k, \quad \text{say} \end{aligned}$$

This can't be the complete solution to the difference equation (1.25), since it contains only a single arbitrary constant. To find the other part of the solution, we try substituting $x_k = k\lambda_3^k$ into the left-hand side of (1.25), producing

$$\begin{aligned} x_{k+2} - ax_{k+1} - bx_k &= (k+2)(\lambda_3)^{k+2} - a(k+1)(\lambda_3)^{k+1} - bk(\lambda_3)^k \\ &= k\lambda_3^k(\lambda_3^2 - a\lambda_3 - b) + \lambda_3^{k+1}(2\lambda_3 - a) \end{aligned} \tag{1.31}$$

The first term within brackets in (1.31) is zero since λ_3 (by definition) satisfies the quadratic equation (1.26), that is to say

$$\begin{aligned} 0 = \lambda^2 - a\lambda - b &\equiv (\lambda - \lambda_3)^2 \\ &\equiv \lambda^2 - 2\lambda_3\lambda + \lambda_3^2 \end{aligned}$$

In this last identity, we see from the terms in λ on both sides that $a = 2\lambda_3$, so the second expression within brackets in (1.31) is also zero. Hence our trial solution works, so the general solution of the difference equation in this case is

$$x_k = c_3\lambda_3^k + c_4 k\lambda_3^k \tag{1.32}$$

where again c_3 and c_4 are constants determined by two given values of x_k.

■ *EXAMPLE 1.9*

The equation

$$x_{k+2} + 6x_{k+1} + 9x_k = 0$$

subject to $x_0 = -1$, $x_1 = 1$ gives rise to the quadratic

$$\lambda^2 + 6\lambda + 9 \equiv (\lambda + 3)^2 = 0$$

Hence $\lambda_3 = -3$, so from (1.32) the solution is

$$x_k = c_3(-3)^k + c_4 k(-3)^k$$

Substituting $k = 0$, $k = 1$ gives

$$\begin{aligned} -1 &= c_3 \\ 1 &= -3c_3 - 3c_4 \end{aligned}$$

so that $c_3 = -1$, $c_4 = \frac{2}{3}$. The solution is therefore

$$x_k = (-3)^k(-1 + \tfrac{2}{3}k), \quad k = 0, 1, 2, \ldots$$

EXERCISE 1.19 Determine the solution of

$$x_{k+2} - 10x_{k+1} + 25x_k = 0$$

subject to $x_1 = 30$, $x_2 = 60$.

EXERCISE 1.20 Obtain the general solution of the equation

$$x_{k+2} - 2ax_{k+1} + x_k = 0, \quad k = 0, 1, 2, 3, \ldots$$

when $|a| < 1$ by setting $a = \cos\ \alpha$ (use the result $e^{i\alpha} = \cos\ \alpha + i\ \sin\ \alpha$).
Determine also the general solution when $a = 1$, and when $a = -1$.

The next level of difficulty is when the right-hand side of the second-order difference equation (1.25) is non-zero. Let's consider just the case when

$$x_{k+2} - ax_{k+1} - bx_k = c, \quad k = 0, 1, 2, \ldots \tag{1.33}$$

where c is a non-zero constant. In the first-order case we found — see (1.21) — that the extra term in the solution was a multiple of c, so by analogy it's natural to try the same thing. Substituting $x_k = pc$, where p is a constant, into (1.33) gives us

$$pc - apc - bpc = c$$

or

$$pc(1 - a - b) = c$$

so that $p = 1/(1 - a - b)$ provided $1 - a - b \neq 0$. The complete solution of (1.33) is then the general solution (whatever it comes out to be) of the equation with $c = 0$, that is the equation

$$x_{k+2} - ax_{k+1} - bx_k = 0 \tag{1.34}$$

plus the extra part $c/(1 - a - b)$. In the standard jargon, (1.34) is called the *homogeneous part* of (1.33), because of the zero right-hand side.

■ *EXAMPLE 1.10*

Consider the equation

$$x_{k+2} + 5x_{k+1} + 6x_k = 10 \tag{1.35}$$

The homogeneous part was solved in Example 1.7. Here $c = 10$, $a = -5$, $b = -6$ so the extra part of the solution due to the term on the right-hand side of (1.35) is

$$10/(1 + 5 + 6) = \tfrac{5}{6}$$

Adding this to the earlier expression in (1.29) gives us the complete solution of (1.35) as

$$x_k = c_1(-2)^k + c_2(-3)^k + \tfrac{5}{6}$$

We need to find out why the method breaks down if $1-a-b=0$. This condition means that the quadratic equation (1.26) has a root $\lambda_1 = 1$, so that the solution of the homogeneous part is

$$x_k = c_1(1)^k + c_2\lambda_2^k$$
$$= c_1 + c_2\lambda_2^k$$

which already contains the 'extra part' $x_k = \text{constant}$. The result given in the following exercise holds in this case.

EXERCISE 1.21 Verify by direct substitution that

$$x_k = ck/(2-a), \quad a \neq 2$$

satisfies the equation

$$x_{k+2} - ax_{k+1} - bx_k = c, \quad 1-a-b=0 \tag{1.36}$$

Similarly, verify that when $a=2$ the complete solution of (1.36) is

$$x_k = c_3 + c_4 k + \tfrac{1}{2}ck^2$$

■ *EXAMPLE 1.11*

Let's solve the equation

$$x_{k+2} + 4x_{k+1} - 5x_k = 9 \tag{1.37}$$

Here $a=-4$, $b=5$, so $1-a-b=0$, and the associated quadratic equation is

$$\lambda^2 + 4\lambda - 5 \equiv (\lambda - 1)(\lambda + 5) = 0$$

which has one root $\lambda_1 = 1$, the other being $\lambda_2 = -5$. Since $c=9$, from Exercise 1.21 the extra term in the solution is $ck/(2-a) = 3k/2$. Hence the general solution of (1.37) is

$$x_k = c_1 + c_2(-5)^k + 3k/2$$

EXERCISE 1.22 Determine the solution of each of the following equations subject to the stated conditions:

(a) $x_{k+2} + 11x_{k+1} + 18x_k = 30$, $x_0 = -1$, $x_1 = 1$

(b) $x_{k+2} - x_k = 1$, $x_0 = 4$, $x_1 = 6$

(c) $4x_{k+2} - 12x_{k+1} + 9x_k = 1$, $x_0 = 0$, $x_1 = 5$

In the next section we'll introduce a slicker way of tackling difference equations. Before doing so, we close this section with another application which is interesting because the variable k does not in this case represent time. However, as the example involves elementary ideas of forces and static equilibrium, those unfamiliar with these concepts can skip directly to the next section.

■ *EXAMPLE 1.12*

A wire (whose mass can be neglected) is tightly stretched between two points as shown in Figure 1.7.

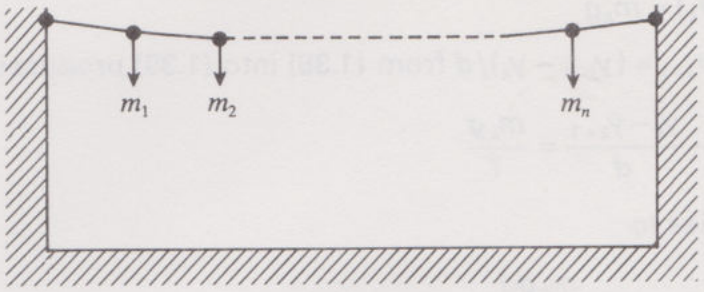


Figure 1.7

Particles having masses $m_1, m_2, \ldots, m_n$ are attached to the wire at equal horizontal distances d apart. In reality the wire will be almost (but cannot be exactly) horizontal, and the situation affecting three neighbouring masses is shown in Figure 1.8.

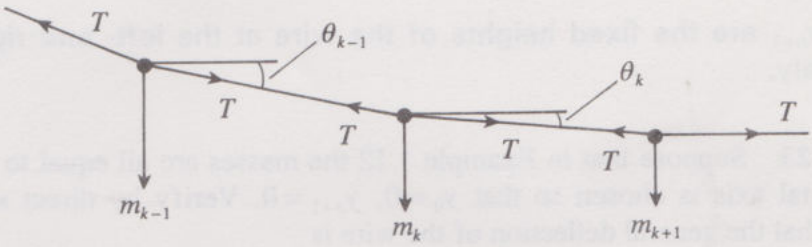


Figure 1.8

The tension in the wire is T, and the angles made with the horizontal have been exaggerated in the figure. Denote by y_k the height of the wire above the ground level for particle m_k, which is at a horizontal distance kd from the left-hand end.

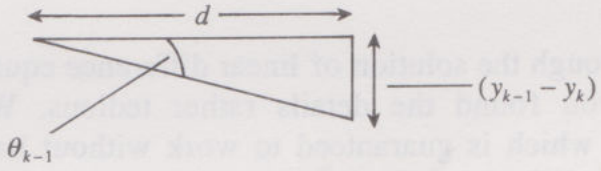


Figure 1.9

We see from Figure 1.9 that

$$\tan \theta_{k-1} = \frac{y_{k-1} - y_k}{d} \tag{1.38}$$

since the portions of wire between each neighbouring pair of weights are assumed straight. Also, since the angles (measured in radians) are small we can assume, to a close approximation, that

$$\tan \theta_{k-1} = \theta_{k-1} = \sin \theta_{k-1}$$

Equating the vertical components of the forces acting on the kth mass gives

$$T \sin \theta_{k-1} - T \sin \theta_k = m_k g$$

so that

$$T(\theta_{k-1} - \theta_k) = m_k g \tag{1.39}$$

Substituting $\theta_{k-1} = (y_{k-1} - y_k)/d$ from (1.38) into (1.39) produces

$$\frac{y_{k-1} - y_k}{d} - \frac{y_k - y_{k+1}}{d} = \frac{m_k g}{T}$$

which simplifies to

$$y_{k+1} - 2y_k + y_{k-1} = \frac{m_k g d}{T}$$

Replacing k by $k+1$ puts this equation into our more familiar form of second-order difference equation:

$$y_{k+2} - 2y_{k+1} + y_k = \frac{m_{k+1} g d}{T} \tag{1.40}$$

and y_0, y_{n+1} are the fixed heights of the wire at the left- and right-hand ends respectively.

EXERCISE 1.23 Suppose that in Example 1.12 the masses are all equal to m, and that the horizontal axis is chosen so that $y_0 = 0$, $y_{n+1} = 0$. Verify by direct substitution into (1.40) that the general deflection of the wire is

$$y_k = \frac{-mgdk(n+1-k)}{2T}, \quad k = 0, 1, 2, \ldots, n+1$$

1.3 THE z-TRANSFORM

As we progressed through the solution of linear difference equations in the previous section, no doubt you found the details rather tedious. What is needed is a foolproof procedure which is guaranteed to work without having to worry about special cases or other difficulties. Such a scheme is provided by the concept of the z-transform, which we'll introduce in this section. The idea of a *transform* is widely used in mathematics, and broadly speaking consists of changing or 'transforming' a problem into a different one which can be solved more easily. For example, the product of two numbers can be found by adding together their logarithms – the problem of multiplication is transformed into the simpler one of addition. Indeed, before cheap pocket calculators became available around 20 years ago, logarithms were widely used for arithmetical work. Admittedly, the simplification to be obtained using the z-transform will not be obvious to you at first, but stay with it!

The solution to any difference equation we are interested in will consist of a *sequence* of values $x_0, x_1, x_2, \ldots$ (assuming that time commences at $k=0$) and we'll use the notation x_k or $\{x_k\}$, with the understanding that $k=0, 1, 2, \ldots$. We define the *z-transform* of this sequence by the series

$$X = x_0 + \frac{x_1}{z} + \frac{x_2}{z^2} + \frac{x_3}{z^3} + \cdots$$

$$= \sum_{k=0}^{\infty} \frac{x_k}{z^k} \tag{1.41}$$

The variable z is simply a parameter which is never assigned a particular value, and the transform X therefore depends on z, often written in the form $X(z)$. Another useful notation is to write $Z(x_k)$ for the z-transform X of the sequence x_k. The sequence and its transform are called a *transform pair*.

■ EXAMPLE 1.13

Suppose that $x_0 = 1$, $x_1 = 3$, $x_2 = -3$, $x_4 = 5, x_k = 0$ for $k \geqslant 5$. Then the z-transform of this sequence is simply

$$X = 1 + \frac{3}{z} - \frac{3}{z^2} + \frac{5}{z^3}$$

In our applications to the solution of linear difference equations the sequence x_k will be infinite. How can the apparently complicated infinite series (1.41) then help us to solve the equations?

There are two steps in answering this. First, for many sequences the series (1.41) can be expressed in a simple form, as we'll now illustrate.

■ EXAMPLE 1.14

(a) The simplest form x_k can take is to be a constant c for all k, so (1.41) gives

$$X = c + \frac{c}{z} + \frac{c}{z^2} + \cdots$$

$$= c\left(1 + \frac{1}{z} + \frac{1}{z^2} + \cdots\right)$$

$$= c\left(1 - \frac{1}{z}\right)^{-1}$$

$$= \frac{cz}{z-1} \tag{1.42}$$

where to get (1.42) we have used the binomial expansion formula (see Problem 1.7):

$$(1+a)^n = 1 + na + \frac{n(n-1)}{2!}a^2 + \frac{n(n-1)(n-2)}{3!}a^3 + \cdots \qquad (1.43)$$

with $a = -1/z$, $n = -1$ (note that $3! = 3 \times 2 \times 1$, $4! = 4 \times 3 \times 2 \times 1$, etc.).

(b) Next, suppose x_k is a *geometric* sequence, that is

$$x_0 = 1,\ x_1 = c,\ x_2 = c^2,\ \ldots,\ x_k = c^k,\ \ldots$$

where c is a constant. From (1.41) we get

$$\begin{aligned} X &= 1 + \frac{c}{z} + \frac{c^2}{z^2} + \cdots \\ &= \left(1 - \frac{c}{z}\right)^{-1} \\ &= \frac{z}{z-c} \end{aligned}$$

where we have again appealed to (1.43), this time with $a = -c/z$ and $n = -1$.

A word is necessary here about the validity of (1.43): strictly speaking, when n is not a positive integer then we must have $|a| < 1$ for the identity to hold. However, we are getting into the mathematical minefield of what is called 'convergence of series'; the aim of this book is not to dwell upon the rigours of mathematics but to convey some ideas and applications. Suffice it to say, then, that for our purposes we can always assume that the parameter z can be made to satisfy such requirements as are necessary for the results to be valid.

EXERCISE 1.24 Determine the z-transforms of the following sequences defined for $k = 0, 1, 2, \ldots$:

(a) $x_k = k$,
(b) $x_k = e^{ck}$, c = constant

Of course, the z-transform of a sequence only needs to be worked out once, and tables of such transforms are readily available, the following (Table 1.1) being a brief sample.

Two points are worth mentioning about Table 1.1. First, it can be used in either direction: we can find the sequence corresponding to a given z-transform by reading from right to left. Secondly, because of the definition in (1.41) we can see that z-transformation is a *linear* operation: that is, according to our earlier discussion, if $Z(x_k) = X$ and $Z(y_k) = Y$ then by the principle of linearity

$$Z(ax_k + by_k) = aX + bY$$

for any constants a and b.

Table 1.1 Some transform pairs

Sequence x_k	Transform $z(x_k)$
c	$\dfrac{cz}{z-1}$
k	$\dfrac{z}{(z-1)^2}$
k^2	$\dfrac{z(z+1)}{(z-1)^3}$
k^3	$\dfrac{z(z^2+4z+1)}{(z-1)^4}$
c^k	$\dfrac{z}{z-c}$
kc^k	$\dfrac{cz}{(z-c)^2}$
$\dfrac{c_2^k - c_1^k}{c_2 - c_1}$	$\dfrac{z}{(z-c_1)(z-c_2)}, \quad c_1 \neq c_2$
$\dfrac{c_2^{k+1} - c_1^{k+1}}{c_2 - c_1}$	$\dfrac{z^2}{(z-c_1)(z-c_2)}, \quad c_1 \neq c_2$

■ EXAMPLE 1.15

By using the appropriate entries in Table 1.1, and the linearity principle, we see that if

$$x_k = 3^k + 5(4)^k \tag{1.44}$$

then its z-transform is

$$\frac{z}{z-3} + \frac{5z}{z-4} = \frac{z(z-4)+5z(z-3)}{(z-3)(z-4)}$$

$$= \frac{z(6z-19)}{z^2-7z+12} \tag{1.45}$$

Alternatively, suppose we are given the z-transform in (1.45) and asked to find the sequence to which it corresponds. We simply factorize the denominator and proceed as follows:

$$\frac{z(6z-19)}{z^2-7z+12} = \frac{6z^2-19z}{(z-3)(z-4)}$$

$$= \frac{6z^2}{(z-3)(z-4)} - \frac{19z}{(z-3)(z-4)}$$

We can obtain the sequence to which each of these transforms corresponds by using the last two entries in Table 1.1. By the principle of linearity, the overall sequence is therefore the sum of these two, that is

$$\begin{aligned} x_k &= \frac{6(4^{k+1}-3^{k+1})}{4-3} - \frac{19(4^k-3^k)}{4-3} \\ &= (-6.3+19)3^k + (6.4-19)4^k \\ &= 3^k + 5(4)^k \end{aligned}$$

which agrees with (1.44).

EXERCISE 1.25 Determine the z-transforms of the following sequences defined for $k=0, 1, 2, 3, \ldots$ by using Table 1.1:

(a) $x_k = 11(-3)^k - 9(-4)^k$

(b) $x_k = 3(2+\mathrm{i})^k - (2-\mathrm{i})^k$

(c) $x_k = (-3)^k(-1+\frac{2}{3}k)$

EXERCISE 1.26 Use Table 1.1 to determine the sequence x_k for each of the following z-transforms:

$$\text{(a)}\ \frac{z}{z^2+11z+18}, \quad \text{(b)}\ \frac{3z^2+z}{z^2-3z+2}, \quad \text{(c)}\ \frac{7z^2-9z}{(z-1)^2}.$$

Our first step towards solving difference equations with the z-transform has been to show how, for a given sequence or transform, we can determine one from the other. You can think of a sequence x_k and its z-transform X as two sides of the same coin: they each represent in different ways the description of some particular process in which we are interested.

We now turn to the second idea which is needed before we can actually proceed to solve difference equations. The key to understanding this is provided by the following argument.

Suppose we shift our sequence to the left, so that the values at $k=0, 1, 2, 3, \ldots$ are now $x_1, x_2, x_3, x_4, \ldots$, as shown in Figure 1.10. Each value has moved one place to the left. We can denote this new sequence by x_{k+1}, where we retain our convention that we start counting at $k=0$. By the definition (1.41), the transform of x_{k+1} is

$$\begin{aligned} \sum_{k=0}^{\infty} \frac{x_{k+1}}{z^k} &= x_1 + \frac{x_2}{z} + \frac{x_3}{z^2} + \frac{x_4}{z^3} + \cdots \\ &= z\left(\frac{x_1}{z} + \frac{x_2}{z^2} + \frac{x_3}{z^3} + \frac{x_4}{z^4} + \cdots\right) \\ &= zX - zx_0 \end{aligned} \tag{1.46}$$

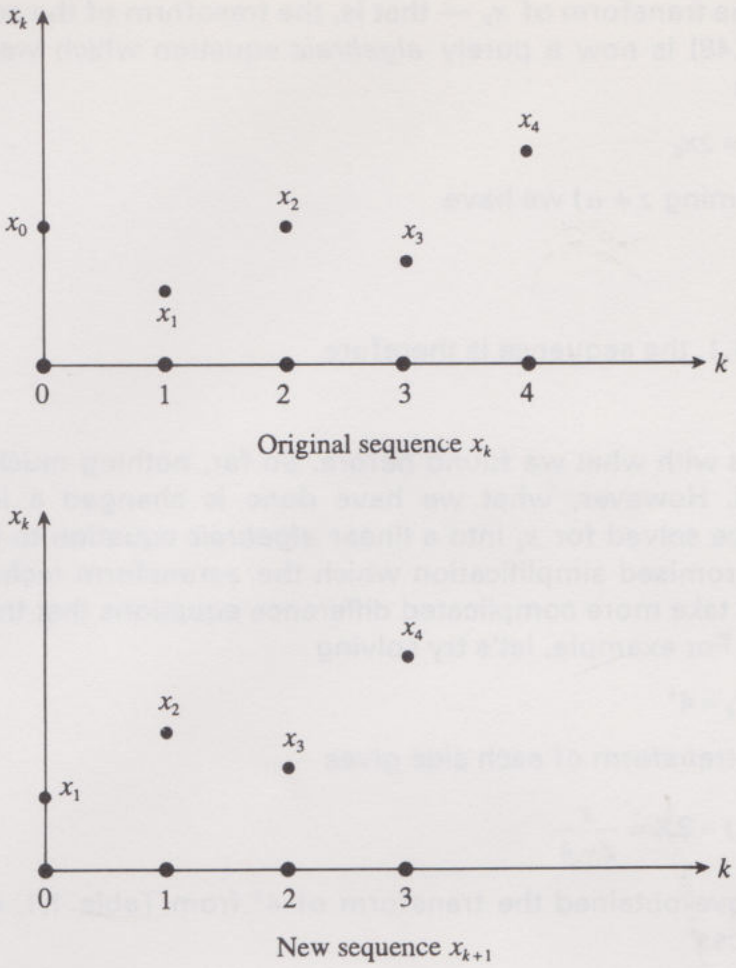


Figure 1.10

where X is the transform of the original sequence x_k, as defined in (1.41). In other words, we have shown that

$$Z(x_{k+1}) = zX - zx_0$$

where x_0 is the *initial value* of x_k.

Thus, apart from the term $-zx_0$, shifting a sequence one place to the left is equivalent to multiplying its transform by z. This is our promised key, which immediately opens the door to solving first-order equations.

■ *EXAMPLE 1.16*

Let's return to the equation

$$x_{k+1} = \alpha x_k, \quad k = 0, 1, 2, 3, \ldots \tag{1.47}$$

first seen in (1.14). The transform of the left-hand side has just been obtained in (1.46), so the transform of the complete equation (1.47) is

$$zX - zx_0 = \alpha X \tag{1.48}$$

where X is the transform of x_k — that is, the transform of the solution of (1.47). However, (1.48) is now a purely *algebraic* equation which we rearrange very easily to give

$$(z-\alpha)X = zx_0$$

so that (assuming $z \neq \alpha$) we have

$$X = \frac{zx_0}{z-\alpha}$$

From Table 1.1, the sequence is therefore

$$x_k = \alpha^k x_0$$

which agrees with what we found before. So far, nothing much seems to have been gained. However, what we have done is changed a linear *difference* equation to be solved for x_k into a linear *algebraic* equation to be solved for X. This is the promised simplification which the z-transform technique produces. It's when we take more complicated difference equations that the benefits really start to flow. For example, let's try solving

$$x_{k+1} - 3x_k = 4^k \tag{1.49}$$

Taking the z-transform of each side gives

$$(zX - zx_0) - 3X = \frac{z}{z-4}$$

where we have obtained the transform of 4^k from Table 1.1. Rearranging the terms produces

$$X = \frac{zx_0}{z-3} + \frac{z}{(z-4)(z-3)}$$

and by again referring to Table 1.1 we immediately obtain

$$x_k = 3^k x_0 + (4^k - 3^k)$$

as the solution of the equation (1.49). Incidentally, obtaining a sequence from its transform is called *inverting* the transform, and x_k is the *inverse* of X.

If the right-hand side in (1.49) is replaced by 3^k, the transformed equation is

$$zX - zx_0 - 3X = \frac{z}{z-3}$$

leading to

$$X = \frac{zx_0}{z-3} + \frac{z}{(z-3)^2}$$

and inverting this transform with the aid of Table 1.1 gives

$$x_k = 3^k x_0 + k3^k$$

The crucial point to realize is that the procedure is purely mechanical — there is no need to try and 'guess' what the extra term is in the solution of the difference equation according to a given right-hand side. Simply do some algebraic

manipulations to get X and then use the table of transforms to find the corresponding x_k.

EXERCISE 1.27 Use the z-transform to solve the two difference equations in Exercise 1.13.

Sometimes the algebra needed is a bit more complicated, as we now see.

■ *EXAMPLE 1.17*

Return to the equation

$$x_{k+1} = 3x_k + k, \quad x_0 = 4$$

previously seen in Exercise 1.16(a). Taking the transform of both sides gives

$$zX - 4z = 3X + \frac{z}{(z-1)^2}$$

so that

$$X = \frac{4z}{z-3} + \frac{z}{(z-1)^2(z-3)} \tag{1.50}$$

The inverse of the first term is $4(3)^k$, but the second term on the right-hand side in (1.50) does not appear in Table 1.1. One way round the difficulty would be to get hold of a book containing a more comprehensive table of transforms! If this is not possible, we need to use what is called 'the method of partial fractions'. This consists of breaking up the nasty term in (1.50) into simpler fractions which *are* listed in the table.

First write

$$\frac{z}{(z-1)^2(z-3)} = z\left[\frac{a}{(z-1)^2} + \frac{b}{z-1} + \frac{c}{z-3}\right] \tag{1.51}$$

where a, b and c are constants to be determined. You should notice two things about (1.51). First, a factor z is kept outside the square brackets because all the transforms in Table 1.1 contain a factor z in the numerator; secondly, it is necessary to include a term over $(z-1)$ as well as one over $(z-1)^2$, as we'll see shortly. The terms inside the square brackets are the *partial fractions* which when added together give the left-hand side, that is

$$\frac{1}{(z-1)^2(z-3)} \equiv \frac{a}{(z-1)^2} + \frac{b}{z-1} + \frac{c}{z-3} \tag{1.52}$$

You may have spotted that (1.52) contains an identity sign instead of an 'equals'. This is because one side is simply a *rearrangement* of the other. We can therefore multiply both sides by $(z-1)^2(z-3)$ so as to remove fractions, giving

$$1 \equiv a(z-3) + b(z-1)(z-3) + c(z-1)^2 \tag{1.53}$$

Remember that (1.53) is an identity, so in particular it holds for any values of z. We choose values which make some terms zero: clearly $z=1$ reduces (1.53) to

$$1 \equiv a(1-3)$$

so that $a=-\frac{1}{2}$, and similarly setting $z=3$ gives

$$1 \equiv c(3-1)^2$$

so that $c=\frac{1}{4}$. Finally, to get the value of b, we look at the term in z^2 in (1.53): this is $(b+c)z^2$ on the right and zero on the left, so $b+c=0$ and hence $b=-c=-\frac{1}{4}$. This shows that we couldn't have started with $b=0$ in (1.51). Putting these values for a, b and c back into (1.51) gives us

$$\frac{z}{(z-1)^2(z-3)} = \frac{-\frac{1}{2}z}{(z-1)^2} - \frac{\frac{1}{4}z}{z-1} + \frac{\frac{1}{4}z}{z-3}$$

Inverting this transform is now possible because all the terms appear in Table 1.1, giving

$$-\tfrac{1}{2}k - \tfrac{1}{4} + \tfrac{1}{4}(3)^k$$

This completes the solution of the original difference equation, and agrees with that given for Exercise 1.16(a).

EXERCISE 1.28

(a) Find the inverse of the transform

$$\frac{z(2z-3)}{(z+2)(z-1)^2}$$

Use this result to solve the difference equation in Exercise 1.16(b).

(b) Show that

$$Z(k^2+3k) = \frac{2z(2z-1)}{(z-1)^3}$$

Use this result to solve the difference equation in Exercise 1.16(c).

Let's now move on to second-order difference equations. Obviously we're going to need the z-transform of the sequence x_{k+2}. Remembering that we start counting at $k=0$, the definition (1.41) gives

$$\begin{aligned} Z(x_{k+2}) &= x_2 + \frac{x_3}{z} + \frac{x_4}{z^2} + \frac{x_5}{z^3} + \cdots \\ &= z^2\left(\frac{x_2}{z^2} + \frac{x_3}{z^3} + \frac{x_4}{z^4} + \cdots\right) \\ &= z^2\left(X - x_0 - \frac{x_1}{z}\right) \\ &= z^2X - z^2x_0 - zx_1 \end{aligned} \tag{1.54}$$

where as usual X is the transform of the original sequence x_k, whose first *two* values

are x_0 and x_1. It is interesting to see that the transform of the *second* difference involves multiplication of X by z^2, just as the transform of the first difference involved zX. This correspondence can be continued, as you may care to try in the following exercise.

EXERCISE 1.29 Show that the z-transform of the sequence x_{k+3}, $k = 0, 1, 2, 3, \ldots$, is

$$z^3X - z^3x_0 - z^2x_1 - zx_2$$

where $Z(x_k) = X$.

We can now apply (1.54) to the solution of second-order difference equations.

■ *EXAMPLE 1.18*

Let's go back to the equation (1.28), namely

$$x_{k+2} + 5x_{k+1} + 6x_k = 0$$

subject to $x_0 = 2$, $x_1 = 3$. Taking the z-transform produces

$$(z^2X - 2z^2 - 3z) + 5(zX - 2z) + 6X = 0$$

which can be rearranged as

$$(z^2 + 5z + 6)X = 2z^2 + 13z$$

so that

$$X = \frac{2z^2}{(z+2)(z+3)} + \frac{13z}{(z+2)(z+3)}$$

Using the last two entries in Table 1.1, the sequence corresponding to X is

$$\begin{aligned} x_k &= 2\,\frac{(-3)^{k+1} - (-2)^{k+1}}{-3-(-2)} + 13\,\frac{(-3)^k - (-2)^k}{-3-(-2)} \\ &= [2(-2)^{k+1} + 13(-2)^k] - [2(-3)^{k+1} + 13(-3)^k] \\ &= (13-4)(-2)^k - (13-6)(-3)^k \\ &= 9(-2)^k - 7(-3)^k \end{aligned}$$

which agrees with the result found earlier in Example 1.7.

EXERCISE 1.30 Repeat Exercise 1.18 using the z-transform and the results in Exercise 1.25(a) and (b).

■ *EXAMPLE 1.19*

Now return to

$$x_{k+2} + 6x_{k+1} + 9x_k = 0, \quad x_0 = -1,\ x_1 = 1$$

considered earlier in Example 1.9. Applying (1.54) gives

$$(z^2X + z^2 - z) + 6(zX + z) + 9X = 0$$

which after some algebraic manipulation becomes

$$X = \frac{-z^2 - 5z}{(z+3)^2} \tag{1.55}$$

We are not quite ready to use Table 1.1, since it doesn't contain a term of the type $z^2/(z-c)^2$. However, if we write (1.55) as

$$\begin{aligned} X &= \frac{-z^2 - 3z - 2z}{(z+3)^2} \\ &= \frac{-z(z+3) - 2z}{(z+3)^2} \\ &= \frac{-z}{z+3} - \frac{2z}{(z+3)^2} \end{aligned}$$

then we can read off from Table 1.1

$$x_k = -(-3)^k + \tfrac{2}{3}k(-3)^k$$

which is the result obtained earlier in Example 1.9.

EXERCISE 1.31 Determine the values of a, b and c such that

$$\frac{z^2 + 5z + 3}{(z+5)(z-1)^2} = \frac{a}{(z-1)^2} + \frac{b}{z-1} + \frac{c}{z+5}$$

Use this result to solve the equation (1.37) subject to the initial conditions $x_0 = 1$, $x_1 = 2$. Compare your answer to that in Example 1.11.

The preceding examples and exercises should have given you the flavour of the z-transform method for solving linear difference equations. It's worth repeating that the essential idea is to transform a difference equation into an algebraic equation, whose solution is then 'inverted' using a table of transforms to give the desired solution of the difference equation. It's important to realize that the solution of the algebraic equation is a routine operation: there is no need to make special provision for awkward cases, as was required in the procedures of the previous section. Indeed, if you have access to a computer algebra package then doing the algebra is completely painless! However, the aim of this section is not to make you an expert at solving difference equations – after all, there are entire books devoted to z-transforms and their applications, which incidentally cover a lot more than merely solving difference equations.

The general concept of 'transforms' is a powerful one in mathematics, and the corresponding method for solving linear *differential* equations, called the *Laplace transform*, has many similarities to the z-transform.

EXERCISE 1.32 Verify the identity

$$\frac{2z}{(z-1)^3} = \frac{z(z+1)}{(z-1)^3} - \frac{z}{(z-1)^2}$$

Use this to find the solution of (1.36) subject to $a = 2$, $x_0 = 0$, $x_1 = 0$. (Compare with the second result in Exercise 1.21.)

EXERCISE 1.33 Return to the problem of the roll of kitchen foil in Exercise 1.17. Solve the difference equation in part (a) using the z-transform. (Hint: find constants a and b such that $X = az/(z-1)^2 + bz(z+1)/(z-1)^3$.)

1.4 MATRIX MODELS

In Examples 1.4 and 1.5 in Section 1.1 we saw how the notation of matrix algebra could be used to describe some discrete time models. We now develop this idea further for some more complicated problems, and show how relevant properties of matrices can be utilized.

■ *EXAMPLE 1.20*

Population models are a particular favourite in this area. Let's consider one which traces the numbers of females in a population of blue whales. The females are divided up into four age groups, and the time period is taken to be 4 years. We use $x_i(k)$, $i = 1, 2, 3, 4$, $k = 0, 1, 2, \ldots$, to denote the number of females in age group i at the beginning of the kth period. Ecological studies have shown that the mortality rate over a 4 year period is 43% for all age groups. The studies also found that the females do not give birth until they are at least 4 years old, and the average numbers of female calves born to a female in group i over a 4 year period are as follows:

Age group i	2	3	4
Age in years	4–7	8–11	12–15
Average number of calves b_i	0.63	1.00	0.90

Age group 1 consists of females aged 0–3 years. After $k+1$ time periods have passed, the number of females $x_1(k+1)$ in this group must be composed of the calves born in the previous time period. From the table we see that $0.63x_2(k)$ calves are born to females in group 2, $1.00x_3(k)$ calves are born to females in group 3 and $0.90x_4(k)$ to females in group 4. We therefore have

$$x_1(k+1) = 0.63x_2(k) + x_3(k) + 0.90x_4(k) \tag{1.56}$$

Now move to the females in group 2. These simply consist of those 57% in group 1 who survive from the previous period, so that

$$x_2(k+1) = 0.57x_1(k)$$

The same argument applies to the next age group, giving

$$x_3(k+1) = 0.57x_2(k)$$

Finally, assuming that no female lives longer than four time periods (16 years) it follows similarly that the oldest age group satisfies

$$x_4(k+1) = 0.57\,x_3(k)$$

We can now write the above four equations in the following combined form:

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \\ x_3(k+1) \\ x_4(k+1) \end{bmatrix} = \begin{bmatrix} 0 & 0.63 & 1 & 0.90 \\ 0.57 & 0 & 0 & 0 \\ 0 & 0.57 & 0 & 0 \\ 0 & 0 & 0.57 & 0 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \\ x_3(k) \\ x_4(k) \end{bmatrix} \tag{1.57}$$

or simply

$$x(k+1) = Ax(k) \tag{1.58}$$

where A is the 4×4 matrix and $x(k)$ is the 4×1 column vector on the right-hand side in (1.57). The elements of A are obtained by picking out the appropriate coefficients in the equations. Specifically, if the ith row of A ($i = 1, 2, 3, 4$) is denoted by a_{i1}, a_{i2}, a_{i3}, a_{i4}, then the ith row of the *product* $Ax(k)$ is

$$a_{i1}x_1(k) + a_{i2}x_2(k) + a_{i3}x_3(k) + a_{i4}x_4(k) \tag{1.59}$$

For example, comparing (1.59) with (1.56) reveals that when $i = 1$ then $a_{11} = 0$, $a_{12} = 0.63$, $a_{13} = 1$, $a_{14} = 0.90$ and these elements comprise the first row of A in (1.57)

The equation (1.58) is called a *matrix difference equation*; in general A can be a square $n \times n$ matrix having n rows and n columns, in which case $x(k)$ is a column vector with n components. When $n = 1$ the equation (1.58) reduces to the scalar equations first seen at the beginning of the chapter in (1.2); the case $n = 2$ appeared in (1.12) and (1.13).

EXERCISE 1.34 In a model of a redwood forest the trees are divided into three age groups:

(1) young trees aged 0 to 200 years;
(2) mature trees aged 200 to 800 years;
(3) old trees aged more than 800 years.

The unit of time k is taken to be 50 years. It is assumed that the trees are uniformly distributed in age throughout each group. Thus, for example, one-quarter of trees in group 1 move into group 2 every 50 years.

In the absence of felling it is reasonable to assume that redwoods die only of old age, and in every 50 year period it is found that one-third of trees in group 3 die. Observations have also shown that in each 50 year period:

each tree in group 1 produces on average 10.25 new trees;
each tree in group 2 produces on average 25 new trees;
each tree in group 3 produces on average 5 new trees.

Let $x_i(k)$, $i=1,2,3$, denote the number of trees in group i at the start of the kth period. Show that

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \\ x_3(k+1) \end{bmatrix} = \begin{bmatrix} 11 & 25 & 5 \\ \frac{1}{4} & \frac{11}{12} & 0 \\ 0 & \frac{1}{12} & \frac{2}{3} \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \\ x_3(k) \end{bmatrix}$$

We can now give the full explanation of what is meant by the product of a matrix A and a column vector $x(k)$. The expression (1.59) is a special case of the general rule. This states that if the element in row i and column j of A is denoted by a_{ij}, and if the n components of $x(k)$ are $x_1(k), x_2(k), \ldots, x_n(k)$, then the product $Ax(k)$ is also a column vector whose ith component is

$$a_{i1}x_1(k) + a_{i2}x_2(k) + \cdots + a_{in}x_n(k), \quad i = 1, 2, \ldots, n \tag{1.60}$$

What (1.60) says is that to get the ith term in the product $Ax(k)$ you take the ith row of A and multiply it term by term with the elements in $x(k)$ – that is, the *first* element in the row of A is multiplied by the *first* element in the column vector, then the second elements are multiplied together, and so on.

EXERCISE 1.35 Evaluate the following products using (1.60):

(a) $$\begin{bmatrix} 1 & -1 & 4 \\ 2 & 3 & 0 \\ 0 & -6 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix}$$

(b) $$\begin{bmatrix} -1 & 2 & 4 & 0 \\ 3 & 1 & -7 & 5 \\ 1 & 0 & 1 & -1 \\ 5 & -1 & 0 & 6 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \\ -3 \\ 4 \end{bmatrix}$$

■ *EXAMPLE 1.21*

In a certain animal population the oldest age attained by females is 15 years. Suppose that the population is divided into three age groups each of duration 5 years, with $x_i(k)$, $i=1,2,3$, denoting the number of females in group i at the beginning of the kth period. Suppose the equation (1.58) in this case has

$$A = \begin{bmatrix} 0 & 3 & 4 \\ \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{2} & 0 \end{bmatrix}, \quad x(k) = \begin{bmatrix} x_1(k) \\ x_2(k) \\ x_3(k) \end{bmatrix}$$

and that initially ($k=0$) there are respectively 1000, 900 and 800 females in each of the three age groups. After 5 years (i.e. one time period) have elapsed the numbers in each age group are given by substituting $k=0$ into (1.58) to give

$$x(1) = Ax(0)$$

Writing this out in full produces

$$\begin{bmatrix} x_1(1) \\ x_2(1) \\ x_3(1) \end{bmatrix} = A \begin{bmatrix} 1000 \\ 900 \\ 800 \end{bmatrix}$$

$$= \begin{bmatrix} 0 \times 1000 + 3 \times 900 + 4 \times 800 \\ \frac{1}{4} \times 1000 + 0 \times 900 + 0 \times 800 \\ 0 \times 1000 + \frac{1}{2} \times 900 + 0 \times 800 \end{bmatrix}$$

$$= \begin{bmatrix} 5900 \\ 250 \\ 450 \end{bmatrix}$$

where we have used the multiplication rule in (1.60) with $n=3$. Thus after 5 years there are 5900 females aged between 0 and 5 years, 250 between 5 and 10 years and 450 between 10 and 15 years.

EXERCISE 1.36 In the preceding example, determine the numbers of females in each of the age groups after a further 5 and 10 years have elapsed by substituting $k=1$ and $k=2$ into (1.58).

The procedure used to solve Example 1.21 and Exercise 1.36 can be followed to solve (1.58) when A is a *general* $n \times n$ matrix. Simply substitute successive values for k into (1.58) starting at zero, just as we did for the scalar equation at the beginning of the chapter. We get

$$k=0: \quad x(1) = Ax(0)$$

$$k=1: \quad x(2) = Ax(1) = A^2x(0)$$

and continuing this process shows that the solution of (1.58) is

$$x(k) = A^k x(0), \quad k = 1, 2, 3, \ldots \tag{1.61}$$

which is the matrix version of (1.3). Remember, however, that when we looked at the scalar solution in more detail in Section 1.2, we had to be careful about what happens to the solution $x(k)$ as k becomes very large. We saw when $n=1$, in which case A in (1.61) is a scalar α, then $\alpha^k \to 0$ as $k \to \infty$ if $|\alpha| < 1$, and $\alpha^k \to \infty$ as $k \to \infty$ if $|\alpha| > 1$. We need to find what replaces this modulus $|\alpha|$ of a real or complex number for the matrix equation case. The easiest situation to deal with is when A is a *diagonal matrix* – that is, the only non-zero entries in A are on the *principal diagonal* (northwest to southeast). For example, a 3×3 diagonal matrix is

$$A = \begin{bmatrix} a_1 & 0 & 0 \\ 0 & a_2 & 0 \\ 0 & 0 & a_3 \end{bmatrix} \tag{1.62}$$

Multiplying A in (1.62) by itself gives

$$A^2 = A.A = \begin{bmatrix} a_1 & 0 & 0 \\ 0 & a_2 & 0 \\ 0 & 0 & a_3 \end{bmatrix}\begin{bmatrix} a_1 & 0 & 0 \\ 0 & a_2 & 0 \\ 0 & 0 & a_3 \end{bmatrix}$$

$$= \begin{bmatrix} a_1^2 & 0 & 0 \\ 0 & a_2^2 & 0 \\ 0 & 0 & a_3^2 \end{bmatrix} \tag{1.63}$$

In obtaining (1.63) we have used the following rule for multiplying matrices, which is really the same as our earlier rule in (1.60) for multiplying together a matrix and a vector.

> *We simply multiply the first matrix with each of the columns of the second matrix in turn.*

For example, with two *general* 3×3 matrices A and B we have

$$AB = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}\begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix}$$

$$\begin{matrix} \uparrow & \uparrow & \uparrow \\ \text{1st} & \text{2nd} & \text{3rd} \\ \text{column} & \text{column} & \text{column} \end{matrix}$$

The *first column* of the product is A times the first column of B, and according to (1.60) this is

$$A\begin{bmatrix} b_{11} \\ b_{21} \\ b_{31} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} \\ a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} \\ a_{31}b_{11} + a_{32}b_{21} + a_{33}b_{31} \end{bmatrix} \tag{1.64}$$

where we have formed the term-by-term product of each of the rows of A with the first column of B.

The second and third columns of AB are found in an exactly similar way to (1.64), using the second and third columns of B.

EXERCISE 1.37 Use the rule (1.64) to evaluate the matrix products

(a) $$\begin{bmatrix} 1 & -1 & 4 \\ 2 & 3 & 0 \\ 0 & -6 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 & 6 \\ 3 & -5 & 11 \\ 2 & 4 & 9 \end{bmatrix}$$

(b) $$\begin{bmatrix} -1 & 2 & 4 & 0 \\ 3 & 1 & -7 & 5 \\ 1 & 0 & 1 & -1 \\ 5 & -1 & 0 & 6 \end{bmatrix} \begin{bmatrix} 2 & 10 & 4 & 3 \\ 1 & 5 & 6 & 11 \\ -3 & -1 & -8 & 0 \\ 4 & 0 & -1 & 9 \end{bmatrix}$$

Compare your results with Exercise 1.35.

Now let's return to the diagonal matrix A in (1.62). Continuing as in (1.63) you can easily check that

$$A^3 = A.A^2 = \begin{bmatrix} a_1^3 & 0 & 0 \\ 0 & a_2^3 & 0 \\ 0 & 0 & a_3^3 \end{bmatrix}$$

and in general

$$A^k = \begin{bmatrix} a_1^k & 0 & 0 \\ 0 & a_2^k & 0 \\ 0 & 0 & a_3^k \end{bmatrix}, \quad k = 2, 3, 4, \ldots$$

With $n = 3$, the solution (1.61) is therefore

$$\begin{bmatrix} x_1(k) \\ x_2(k) \\ x_3(k) \end{bmatrix} = A^k \begin{bmatrix} x_1(0) \\ x_2(0) \\ x_3(0) \end{bmatrix}$$

$$= \begin{bmatrix} a_1^k x_1(0) \\ a_2^k x_2(0) \\ a_3^k x_3(0) \end{bmatrix} \tag{1.65}$$

again using the rule (1.64). Equating components on each side of (1.65) shows that

$$x_i(k) = a_i^k x_i(0), \quad i = 1, 2, 3$$

This is now in exactly the same form as the solution (1.3) for the scalar case, and we can therefore conclude that as $k \to \infty$

$$\begin{aligned} &x_i(k) \to 0, \quad \text{if } |a_i| < 1 \\ &x_i(k) \to \infty, \quad \text{if } |a_i| > 1 \end{aligned} \tag{1.66}$$

Although we have only worked with the 3×3 matrix in (1.62), you should be able to see that the argument used to derive (1.66) holds for *any* value of n when A is a diagonal matrix with diagonal entries $a_1, a_2, \ldots, a_n$.

To handle the situation when A is an arbitrary matrix, not a diagonal form, it's necessary to delve into a bit of matrix theory. As usual in this book, we'll keep the treatment informal – for full details and a mathematically rigorous coverage you'll have to go to an appropriate book in the reading list at the end of the chapter. The basic idea is to change the coordinates in our difference equation (1.58) from $x_1, x_2, \ldots, x_n$ to $y_1, y_2, \ldots, y_n$ according to a linear relationship – that is, each x_i is a *linear combination* of y, which means, for example, that

$$x_1 = t_{11}y_1 + t_{12}y_2 + \cdots + t_{1n}y_n \tag{1.67}$$

where the ts are constants (notice that for simplicity of notation we have temporarily suppressed the dependence of the xs and the ys on the variable k). Putting together the expressions like (1.67) for each of the xs, we get $x = Ty$ where T is the $n \times n$ matrix whose element in row i, column j is t_{ij}, and x and y are column vectors with components $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ respectively. To express y in terms of x, we would somehow like to 'divide' by T, to obtain $y = x \div T$. The way we do this is to define the *inverse* T^{-1} of T. This matrix satisfies the conditions

$$TT^{-1} = T^{-1}T = I$$

where I is the $n \times n$ *unit matrix*, which has diagonal form like (1.62) with ones all along the diagonal. You can easily check using our multiplication rule that

$$XI = IX = X$$

for any $n \times n$ matrix X. This accounts for the name 'unit matrix', since I behaves for matrices in the same way that the unit 1 behaves for scalars. Similarly, the name 'inverse' of T is used because T^{-1} plays exactly the same role for matrices as does the inverse t^{-1} $(= 1/t)$ of a scalar t. We can therefore rearrange $x = Ty$ to obtain

$$y = T^{-1}x$$

(be careful to write the inverse on the correct side: xT^{-1} does not make sense).

EXERCISE 1.38 Verify in each of the following cases that $AB = BA = I$, where I is the 2×2 unit matrix. This shows that $B = A^{-1}$.

(a) $A = \begin{bmatrix} 3 & 1 \\ 2 & 1 \end{bmatrix}$, $B = \begin{bmatrix} 1 & -1 \\ -2 & 3 \end{bmatrix}$

(*b*) $A = \begin{bmatrix} 4 & -1 \\ 3 & 2 \end{bmatrix}$, $B = \frac{1}{11}\begin{bmatrix} 2 & 1 \\ -3 & 4 \end{bmatrix} = \begin{bmatrix} \frac{2}{11} & \frac{1}{11} \\ -\frac{3}{11} & \frac{4}{11} \end{bmatrix}$

Let's now see what happens to

$$x(k+1) = Ax(k)$$

when we change the variables from x to y. Firstly we have

$$x(k+1) = ATy(k)$$

and then using the argument above gives

$$\begin{aligned} y(k+1) &= T^{-1}x(k+1) \\ &= T^{-1}ATy(k) \end{aligned} \tag{1.68}$$

This difference equation in the new variables $y_1(k), y_2(k), \ldots, y_n(k)$ can therefore be written

$$y(k+1) = Dy(k), \quad k = 0, 1, 2, \ldots \tag{1.69}$$

where $D = T^{-1}AT$. The crucial fact which is relevant for us is that in most cases of practical interest it's possible to find a matrix T which possesses an inverse T^{-1} and is such that D is *diagonal*, that is

$$T^{-1}AT = D = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_n \end{bmatrix} \tag{1.70}$$

Although T is not unique, the non-zero elements along the diagonal of D *are* unique for a given matrix A, although for different Ts they may come out in different orders. These numbers $\lambda_1, \lambda_2, \ldots, \lambda_n$ are therefore specific to a given matrix A and are called the *eigenvalues* of A. Words involving 'eigen-' are rather ugly Anglo-German combinations but are widely used – the term *characteristic roots* of A is an alternative expression. Since $y(k)$ satisfies (1.69), where D is the diagonal matrix in (1.70), we know from (1.66) that the behaviour of $y(k)$ as $k \to \infty$ is

$$y_i(k) \to 0, \quad \text{if } |\lambda_i| < 1$$

$$y_i(k) \to \infty, \quad \text{if } |\lambda_i| > 1$$

Recall that each component $x_i(k)$ is a linear combination of all the components of $y(k)$, as displayed in (1.67) for $x_1(k)$. It therefore follows that as $k \to \infty$ we must have *all* the $y_i(k) \to 0$ in order to make each $x_i(k) \to 0$. That is, as $k \to \infty$ we have

$$x(k) \to 0 \quad \text{provided } \textit{all } |\lambda_i| < 1 \tag{1.71}$$

Similarly, if *all* $|\lambda_i| > 1$ then $x(k) \to \infty$ as $k \to \infty$. You might like to think about what happens if some of the λ_i have modulus less than one, and the remainder have modulus greater than one.

In order to apply the result (1.71), we need to be able to compute these mysterious quantities called eigenvalues for any given matrix A. This is a subject which can and does fill whole books! Nevertheless, let's see how far we can get without becoming too technical. The key equation is (1.70), called *diagonalization* of A, whereby A is converted into a diagonal matrix D according to

$$T^{-1}AT = D$$

Multiplying this (on the left) by T gives

$$AT = TD \tag{1.72}$$

since $TT^{-1} = I$ and $IA = A$. Suppose the matrix T has columns $t_1, t_2, \ldots, t_n$. For simplicity consider $n = 2$, so (1.72) becomes

$$A\begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix} = \begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix}\begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

$$\begin{matrix} \uparrow & \uparrow & \uparrow & \uparrow \\ t_1 & t_2 & t_1 & t_2 \end{matrix}$$

$$= \begin{bmatrix} \lambda_1 t_{11} & \lambda_2 t_{12} \\ \lambda_1 t_{21} & \lambda_2 t_{22} \end{bmatrix}$$

using the multiplication rule given earlier in (1.60). Equate the first columns on each side of this equation to get

$$A\begin{bmatrix} t_{11} \\ t_{21} \end{bmatrix} = \begin{bmatrix} \lambda_1 t_{11} \\ \lambda_1 t_{21} \end{bmatrix}$$

or

$$At_1 = \lambda_1 t_1$$

using the fact that multiplying a vector by a scalar means that each element is multiplied by the scalar. Similarly using the second columns gives $At_2 = \lambda_2 t_2$. Generalizing this approach, when A is an $n \times n$ matrix we get n equations of the form

$$Az = \lambda z \tag{1.73}$$

to be solved for the scalar λ and the column vector z. In general there will be n such values of λ, called *eigenvalues*, and for each value there will be a corresponding vector z called an *eigenvector* (these, however, are not unique).

■ *EXAMPLE 1.22*

Let's solve (1.73) for $n = 2$ in the following case:

$$\begin{bmatrix} 1 & 3 \\ 2 & 2 \end{bmatrix}\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \lambda\begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$$

$$\begin{matrix} A & & z & & z \end{matrix}$$

Using the multiplication rule (1.60), this becomes

$$\begin{bmatrix} z_1 + 3z_2 \\ 2z_1 + 2z_2 \end{bmatrix} = \begin{bmatrix} \lambda z_1 \\ \lambda z_2 \end{bmatrix}$$

Equating terms on each side gives us

$$z_1 + 3z_2 = \lambda z_1, \quad 2z_1 + 2z_2 = \lambda z_2$$

or, after rearrangement,

$$\begin{aligned} (\lambda - 1)z_1 - 3z_2 &= 0 \\ -2z_1 + (\lambda - 2)z_2 &= 0 \end{aligned} \tag{1.74}$$

We have *two* simultaneous equations which involve *three* unknowns λ, z_1 and z_2 – we'll see that this means the solution is not unique. Let's eliminate z_2 by multiplying the first equation in (1.74) by $(\lambda - 2)$ and the second equation by 3:

$$\begin{aligned} (\lambda - 2)(\lambda - 1)z_1 - 3(\lambda - 2)z_2 &= 0 \\ -6z_1 + 3(\lambda - 2)z_2 &= 0 \end{aligned}$$

Now add these equations to obtain

$$[(\lambda - 2)(\lambda - 1) - 6]z_1 = 0 \tag{1.75}$$

We cannot have $z_1 = 0$, for if this were the case the first equation in (1.74) would give $z_2 = 0$ as well – completely uninteresting! We therefore conclude that the content of the square bracket in (1.75) is zero, that is

$$\lambda^2 - 3\lambda - 4 = 0 \tag{1.76}$$

or

$$(\lambda + 1)(\lambda - 4) = 0$$

This shows that the two eigenvalues of A are the roots of the quadratic equation (1.76), namely $\lambda_1 = -1$, $\lambda_2 = 4$. We now go back to (1.74) and solve the equations for z_1 and z_2 using each of these two values of λ. With $\lambda = -1$, (1.74) becomes

$$\begin{aligned} -2z_1 - 3z_2 &= 0 \\ -2z_1 - 3z_2 &= 0 \end{aligned} \tag{1.77}$$

You can see that a convenient solution of (1.77) is $z_1 = 3, z_2 = -2$. However, as mentioned above, the solution of (1.77) is not unique – clearly $z_1 = 3p$, $z_2 = -2p$ is also a solution for any scalar p.

When $\lambda = 4$ the equations (1.74) become

$$\begin{aligned} 3z_1 - 3z_2 &= 0 \\ -2z_1 + 2z_2 &= 0 \end{aligned} \tag{1.78}$$

for which a convenient solution is $z_1 = 1$, $z_2 = 1$. We have therefore found eigenvectors

$$\begin{bmatrix} 3 \\ -2 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

corresponding to the eigenvalues λ_1, λ_2 respectively. These eigenvectors are the columns of the matrix T in (1.70), which is therefore

$$T = \begin{bmatrix} 3 & 1 \\ -2 & 1 \end{bmatrix} \tag{1.79}$$

To complete (1.70) we need the inverse T^{-1}, and it's useful to quote here the formula for the inverse of *any* 2×2 matrix:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \tag{1.80}$$

provided $ad - bc \neq 0$. To verify that (1.80) is correct, simply multiply the matrix and its inverse together, using the rule in (1.64):

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}\frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

$$= \frac{1}{ad-bc}\begin{bmatrix} ad-bc & -ab+ba \\ cd-dc & -cb+ad \end{bmatrix}$$

$$= \frac{1}{ad-bc}\begin{bmatrix} ad-bc & 0 \\ 0 & -cb+ad \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I$$

as required. Two examples of (1.80) were given in Exercise 1.38. We can now apply (1.80) to (1.79) to get

$$T^{-1} = \frac{1}{3-(-2)}\begin{bmatrix} 1 & -1 \\ 2 & 3 \end{bmatrix}$$

$$= \frac{1}{5}\begin{bmatrix} 1 & -1 \\ 2 & 3 \end{bmatrix}$$

When T in (1.79) and its inverse above are put into (1.70) you should verify by working out the products that we do indeed get the diagonal matrix D:

$$T^{-1}AT = \begin{bmatrix} -1 & 0 \\ 0 & 4 \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \tag{1.81}$$

Having gone through this example in some detail, we can now expand on a few points which will enable us to solve any 2×2 example rather more succinctly. First, notice the condition that (1.80) must not have a zero denominator: the quantity $ad - bc$ is called the *determinant* of the 2×2 matrix

$$X = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

since it *determines* whether this matrix has an inverse, and is written $\det X$ or $|X|$. Secondly, the equation (1.73) can be written as

$$(\lambda I - A)z = 0 \tag{1.82}$$

since $\lambda z = \lambda I z$. Moreover, the condition that $z \neq 0$ can be shown to require that

$$\det(\lambda I - A) = 0 \tag{1.83}$$

As an illustration, consider the 2×2 matrix A in Example 1.22 for which

$$\lambda I - A = \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 1 & 3 \\ 2 & 2 \end{bmatrix}$$

$$= \begin{bmatrix} \lambda - 1 & -3 \\ -2 & \lambda - 2 \end{bmatrix}$$

Hence (1.83) gives

$$0 = (\lambda - 1)(\lambda - 2) - (-3)(-2)$$
$$= \lambda^2 - 3\lambda - 4$$

which is precisely what we found before in (1.76). The eigenvalues of A can therefore be found by solving (1.83), which is called the *characteristic equation* of A (the polynomial itself is called the *characteristic polynomial* of A). When A is a general $n \times n$ matrix (1.83) still applies, and the characteristic polynomial has degree n. However, we shan't go into details of how to work out the determinant of a general square matrix, although we shall consider shortly the cases $n = 3$ and $n = 4$.

Finally, there's no need to worry about the non-unique solution for an eigenvector – whatever you choose, the diagonal expression $T^{-1}AT$ will still come out right. For example, suppose that in Example 1.22 we had chosen $z_1 = 6$, $z_2 = -4$ as a solution of (1.77), and $z_1 = 3$, $z_2 = 3$ as a solution of (1.78). The new matrix T would then be

$$T = \begin{bmatrix} 6 & 3 \\ -4 & 3 \end{bmatrix}$$

with inverse given by (1.80) as

$$T^{-1} = \frac{1}{18 - (-12)} \begin{bmatrix} 3 & -3 \\ 4 & 6 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{10} & -\frac{1}{10} \\ \frac{2}{15} & \frac{1}{5} \end{bmatrix}$$

You should verify by multiplying out the product that we still get

$$T^{-1}AT = \begin{bmatrix} -1 & 0 \\ 0 & 4 \end{bmatrix}$$

as in (1.81).

EXERCISE 1.39 Determine the eigenvalues and associated eigenvectors for the matrix

$$A = \begin{bmatrix} 1 & 3 \\ 3 & 1 \end{bmatrix}$$

Hence obtain T and T^{-1} in (1.70), and verify that $T^{-1}AT = D$.

Remember that an eigenvector z linked to a *particular* eigenvalue λ satisfies (1.73), namely $Az = \lambda z$. Multiplying both sides by A produces

$$A^2z = \lambda Az = \lambda(\lambda z) = \lambda^2 z$$

Continuing this process shows that $A^3z = \lambda^3 z$ and in general

$$A^k z = \lambda^k z \tag{1.84}$$

for any positive integer k. Recall also that the solution of the matrix difference equation (1.58), namely

$$x(k+1) = Ax(k), \quad k = 1, 2, 3, \ldots$$

was obtained in (1.61) as $x(k) = A^k x(0)$. It's interesting to realize that if the initial vector $x(0)$ is an eigenvector z of A, then using (1.84) shows that the solution is

$$x(k) = A^k z = \lambda^k z$$

Moreover, if this particular eigenvalue has the value 1 then $\lambda^k = 1$, and the solution is simply $x(k) = z$: that is, the vector $x(k)$ remains at its initial value $x(0) = z$ for all subsequent time. This is called an *equilibrium* situation.

■ EXAMPLE 1.23

A local car rental company has two offices in neighbouring cities B and L. It is known on the basis of past experience that on a monthly basis 40% of rentals from the office in B are returned there and 60% are one-way rentals which are dropped off in L. Similarly, 70% of rentals from the office in L are returned there, whereas 30% are dropped off in B. The company operates a fleet of 90 cars, and would like to know the numbers which should be kept at each city depot.

Let x_k, y_k denote the number of cars at the depots in B and L respectively at the beginning of month k, for $k = 0, 1, 2, \ldots$. One month later the cars at B consist of those returned there during the previous month (namely 40% of x_k), together with those dropped off on a one-way rental from L (i.e. 30% of y_k) so we have

$$x_{k+1} = 0.4x_k + 0.3y_k, \quad k = 0, 1, 2, \ldots$$

Similarly, for the depot at L

$$y_{k+1} = \underbrace{0.6x_k}_{\text{cars from B}} + \underbrace{0.7y_k}_{\text{cars returned to L}}$$

We can write these equations in the combined form

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \underbrace{\begin{bmatrix} 0.4 & 0.3 \\ 0.6 & 0.7 \end{bmatrix}}_{A} \begin{bmatrix} x_k \\ y_k \end{bmatrix}, \quad k = 0, 1, 2, \ldots \tag{1.85}$$

which is precisely our standard equation $X(k+1) = AX(k)$ with

$$X(k) = \begin{bmatrix} x_k \\ y_k \end{bmatrix}$$

Let's find the eigenvalues of A using (1.83), which gives

$$0 = \det(\lambda I - A) = \det\begin{bmatrix} \lambda - 0.4 & -0.3 \\ -0.6 & \lambda - 0.7 \end{bmatrix}$$
$$= (\lambda - 0.4)(\lambda - 0.7) - (-0.3)(-0.6)$$
$$= \lambda^2 - 1.1\lambda + 0.1$$
$$= (\lambda - 1)(\lambda - 0.1)$$

Hence A has eigenvalues $\lambda_1 = 1$, $\lambda_2 = 0.1$. An eigenvector for $\lambda = 1$ is found by solving (1.82), which becomes

$$\underbrace{\begin{bmatrix} 0.6 & -0.3 \\ -0.6 & 0.3 \end{bmatrix}}_{(\lambda_1 I - A)} \underbrace{\begin{bmatrix} z_1 \\ z_2 \end{bmatrix}}_{z} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

On multiplying this out, we get the same equation twice, namely

$$0.6z_1 - 0.3z_2 = 0$$

so we can take $z_1 = p$, $z_2 = 2p$ for any non-zero scalar p as the components of an eigenvector for the eigenvalue $\lambda = 1$. We therefore conclude from our discussion above that if we start with $x_0 = p$, $y_0 = 2p$ then $x_k = p$, $y_k = 2p$ for all subsequent time, $k = 1, 2, 3, \ldots$. Since the total number of cars is $p + 2p = 90$ we have $p = 30$, so the company should begin with 30 cars at B and 60 at L; these numbers will then be the same at the beginning of every subsequent month (assuming that the previous pattern of customer usage does not alter) and this is an equilibrium situation.

EXERCISE 1.40 Verify by substituting $x_0 = p$, $y_0 = 2p$ into (1.85) that $x_k = p$, $y_k = 2p$ for all values of $k \geqslant 1$.

EXERCISE 1.41 If the matrix A in (1.85) is changed to

$$\begin{bmatrix} \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{2}{3} \end{bmatrix}$$

show that it still has an eigenvalue equal to 1.

Determine the numbers of cars to be kept at B and L for the equilibrium situation in this case.

EXERCISE 1.42 Show that any matrix of the form

$$\begin{bmatrix} \alpha & \beta \\ 1 - \alpha & 1 - \beta \end{bmatrix}$$

where α and β are positive scalars, has an eigenvalue equal to 1.

We've seen that when $n = 2$ the determinant in (1.83) gives a quadratic equation to be solved for λ. It turns out that when $n = 3$ we get a cubic equation in λ, when $n = 4$ a fourth-degree equation, and so on. Let's consider the barest outline of how a

3×3 determinant can be worked out. Let x_{ij} denote the element in row i, column j of an arbitrary matrix X. We already know that for $n = 2$

$$\det X = \begin{vmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{vmatrix} = x_{11}x_{22} - x_{12}x_{21}$$

When $n = 3$, $\det X$ is defined in terms of three 2×2 determinants:

$$\det X = \begin{vmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{vmatrix}$$

$$= x_{11}X_{11} - x_{12}X_{12} + x_{13}X_{13} \tag{1.86}$$

where

$$X_{11} = \begin{vmatrix} x_{22} & x_{23} \\ x_{32} & x_{33} \end{vmatrix}, \quad X_{12} = \begin{vmatrix} x_{21} & x_{23} \\ x_{31} & x_{33} \end{vmatrix}, \quad X_{13} = \begin{vmatrix} x_{21} & x_{22} \\ x_{31} & x_{32} \end{vmatrix} \tag{1.87}$$

To form: X_{11} we delete row 1, column 1 in $\det X$
X_{12} we delete row 1, column 2 in $\det X$
X_{13} we delete row 1, column 3 in $\det X$.

Analogous formulae can be given for $n \geqslant 4$, but for further explanations and proofs you'll have to consult an appropriate book from the list at the end of the chapter. In fact, for $n \geqslant 4$ evaluating determinants using formulae like (1.86) is not recommended – a much better method relies on what is called 'gaussian elimination' (see Section 3.5, Chapter 3).

■ *EXAMPLE 1.24*

We'll find eigenvalues and corresponding eigenvectors for the matrix

$$A = \begin{bmatrix} 0 & -1 & 1 \\ 2 & 3 & 3 \\ -2 & 1 & 1 \end{bmatrix}$$

Write X for the matrix

$$\lambda I - A = \begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix} - A$$

We evaluate $\det X = \det(\lambda I - A)$ using (1.86) as follows:

$$\det X = \begin{vmatrix} \lambda & 1 & -1 \\ -2 & \lambda - 3 & -3 \\ 2 & -1 & \lambda - 1 \end{vmatrix}$$

$$= \lambda \underbrace{\begin{vmatrix} \lambda - 3 & -3 \\ -1 & \lambda - 1 \end{vmatrix}}_{X_{11}} - \underbrace{\begin{vmatrix} -2 & -3 \\ 2 & \lambda - 1 \end{vmatrix}}_{X_{12}} + (-1) \underbrace{\begin{vmatrix} -2 & \lambda - 3 \\ 2 & -1 \end{vmatrix}}_{X_{13}}$$

$$= \lambda[(\lambda - 3)(\lambda - 1) - (-3)(-1)] - [-2(\lambda - 1) - (-3)2] - [(-2)(-1) - 2(\lambda - 3)]$$

$$= \lambda(\lambda^2 - 4\lambda) - (-2\lambda + 8) - (-2\lambda - 8)$$

$$= \lambda^3 - 4\lambda^2 + 4\lambda - 16 \qquad (1.88)$$

By trying simple values you can check that $\lambda = 4$ is a root of this polynomial, so $\det X$ has a factor $\lambda - 4$. Dividing the polynomial in (1.88) by this factor produces

$$\det(\lambda I - A) = (\lambda - 4)(\lambda^2 + 4)$$
$$= (\lambda - 4)(\lambda + 2\mathrm{i})(\lambda - 2\mathrm{i})$$

Hence the eigenvalues of A are $\lambda_1 = 4$, $\lambda_2 = 2\mathrm{i}$, $\lambda_3 = -2\mathrm{i}$. If A is a real matrix, as is almost invariably the case in practical applications, then any complex eigenvalues *always* occur in complex conjugate pairs in the form $\alpha \pm \mathrm{i}\beta$.

To find an eigenvector corresponding to $\lambda_1 = 4$, we solve the equation (1.82), which here is

$$\underbrace{\begin{bmatrix} 4 & 1 & -1 \\ -2 & 1 & -3 \\ 2 & -1 & 3 \end{bmatrix}}_{\lambda I - A} \underbrace{\begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}}_{z} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Working out this product using the multiplication rule (1.60), we get

$$\begin{aligned} 4z_1 + z_2 - z_3 &= 0 \\ -2z_1 + z_2 - 3z_3 &= 0 \\ 2z_1 - z_2 + 3z_3 &= 0 \end{aligned}$$

Again the solution of these equations is not unique, since you can see that the last two equations are identical. Subtract the second equation from the first to get

$$6z_1 + 2z_3 = 0$$

for which a simple solution is $z_1 = 1$, $z_3 = -3$. From the first equation

$$z_2 = -4z_1 + z_3 = -4 - 3 = -7$$

so an eigenvector for λ_1 is

$$\begin{bmatrix} 1 \\ -7 \\ -3 \end{bmatrix}$$

EXERCISE 1.43 Determine eigenvectors corresponding to λ_2 and λ_3 for the matrix A in Example 1.24.

We will not develop any further the question of computation of eigenvalues of a matrix. Not only is powerful software available for this purpose on microcomputers, but also on some hand-held graphics calculators. It is interesting, however, to return to our population models. If you go back to Examples 1.20 and 1.21 at the beginning of this section you will see that in each case the matrix A has the form

$$L = \begin{bmatrix} a_1 & a_2 & a_3 & \dots & a_{n-1} & a_n \\ b_1 & 0 & 0 & \dots & 0 & 0 \\ 0 & b_2 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \dots & b_{n-1} & 0 \end{bmatrix} \tag{1.89}$$

where $a_i \geqslant 0$ and $1 \geqslant b_i > 0$ for $i = 1, 2, 3, \ldots$.

The matrix (1.89) is called a *Leslie* matrix, and we can generalize our models of female populations given in Examples 1.20 and 1.21 in the following way. Divide up the population into n age groups of equal duration. If the maximum age attained by any females is K years, then the time period has length K/n years. For example, in our blue whale model in Example 1.20 we had $K = 16$, $n = 4$ and the length of the time period was 4 years. Let $x_i(k)$ denote the number of females in group i at time k, and interpret the parameters in the matrix L in (1.89) as:

a_i = the average number of daughters born to a female during the time she is in the ith age group

b_i = the fraction of females in the ith age group which survive to pass into the $(i+1)$th group

Using exactly the same argument as in our two particular cases in Examples 1.20 and 1.21 we can construct the equations which describe the way the population behaves. First, $x_1(k+1)$ is equal to the total of the daughters born to females in all the age groups over the previous time period, so that

$$x_1(k+1) = a_1 x_1(k) + a_2 x_2(k) + \cdots + a_n x_n(k) \tag{1.90}$$

Secondly, the number of females in group $i+1$ at time $k+1$ is equal to the proportion of group i surviving from the previous time period, so that

$$x_{i+1}(k+1) = b_i x_i(k), \quad i = 1, 2, 3, \ldots, n-1 \tag{1.91}$$

Combining together the n equations represented by (1.90) and (1.91) gives

$$X(k+1) = LX(k) \tag{1.92}$$

where $X(k)$ has components $x_1(k), x_2(k), \ldots, x_n(k)$, and L is the Leslie matrix in (1.89).

The Leslie matrix has some interesting properties which we can use to investigate the general behaviour of the solution of the population equation (1.92). Of course, since this equation has our usual standard form we know from (1.61) that the solution of (1.92) is

$$X(k) = L^k X(0), \quad k = 1, 2, 3, \ldots \tag{1.93}$$

We have seen that the expression (1.93) will depend upon the eigenvalues and eigenvectors of L; an important fact which can be proved is that any matrix L in the form (1.84) always has at least one positive eigenvalue, which we shall denote by λ_1, and this is *unique* (i.e. it occurs only once as a root of the characteristic equation of L).

■ *EXAMPLE 1.25*

Consider the case $n = 3$, when

$$L = \begin{bmatrix} a_1 & a_2 & a_3 \\ b_1 & 0 & 0 \\ 0 & b_2 & 0 \end{bmatrix} \tag{1.94}$$

The characteristic equation of L is

$$\begin{aligned} 0 &= \det(\lambda I - L) \\ &= \begin{bmatrix} \lambda - a_1 & -a_2 & -a_3 \\ -b_1 & \lambda & 0 \\ 0 & -b_2 & \lambda \end{bmatrix} \\ &= (\lambda - a_1)\begin{bmatrix} \lambda & 0 \\ -b_2 & \lambda \end{bmatrix} - (-a_2)\begin{bmatrix} -b_1 & 0 \\ 0 & \lambda \end{bmatrix} - a_3\begin{bmatrix} -b_1 & \lambda \\ 0 & -b_2 \end{bmatrix} \\ &= (\lambda - a_1)\lambda^2 + a_2(-b_1\lambda) - a_3 b_1 b_2 \\ &= \lambda^3 - a_1\lambda^2 - a_2 b_1 \lambda - a_3 b_1 b_2 \end{aligned} \tag{1.95}$$

where to obtain (1.95) we again used the expressions (1.86) and (1.87). Let λ_1 be the positive eigenvalue of L in (1.94), and consider the product

$$L\begin{bmatrix} 1 \\ b_1/\lambda_1 \\ b_1 b_2/\lambda_1^2 \end{bmatrix} = \begin{bmatrix} a_1 + a_2 b_1/\lambda_1 + a_3 b_1 b_2/\lambda_1^2 \\ b_1 \\ b_1 b_2/\lambda_1 \end{bmatrix} \tag{1.96}$$

Since λ_1 satisfies the equation (1.95), we can write

$$\lambda_1^3 = a_1\lambda_1^2 + a_2 b_1 \lambda_1 + a_3 b_1 b_2$$

and dividing both sides of this by λ_1^2 shows that the first element on the

right-hand side of (1.96) is just λ_1. The equation (1.96) therefore reduces to

$$Lu = \lambda_1 u \tag{1.97}$$

where

$$u = \begin{bmatrix} 1 \\ b_1/\lambda_1 \\ b_1 b_2/\lambda_1^2 \end{bmatrix} \tag{1.98}$$

showing that u is an eigenvector corresponding to λ_1.

EXERCISE 1.44 Verify using (1.95) that the Leslie matrix

$$L = \begin{bmatrix} 0 & 7 & 6 \\ \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{2} & 0 \end{bmatrix} \tag{1.99}$$

has an eigenvalue $\lambda_1 = \frac{3}{2}$. Determine a corresponding eigenvector using (1.98), and verify that it satisfies (1.97). Find also the other two eigenvalues of L.

We now list some properties of the general $n \times n$ Leslie matrix L in (1.89):

(i)
$$\det(\lambda I - L) = \lambda^n - a_1\lambda^{n-1} - a_2 b_1 \lambda^{n-2} - a_3 b_1 b_2 \lambda^{n-3} \ldots - a_n b_1 b_2 \ldots b_{n-1}$$
The case $n = 3$ is given in (1.95).

(ii) There is a unique positive eigenvalue λ_1 with corresponding eigenvector

$$u = \begin{bmatrix} 1 \\ b_1/\lambda_1 \\ b_1 b_2/\lambda_1^2 \\ b_1 b_2 b_3/\lambda_1^3 \\ \vdots \\ b_1 b_2 \ldots b_{n-1}/\lambda_1^{n-1} \end{bmatrix} \tag{1.100}$$

The case $n = 3$ is given in (1.98).

(iii) If λ_i is *any* other eigenvalue (real or complex) of L then $|\lambda_i| \leqslant \lambda_1$, and λ_1 is called the *dominant* eigenvalue of L. Furthermore, provided that in the first row of L there are two successive entries a_i and a_{i+1} which are both non-zero, then λ_1 is *strictly dominant*, which means $|\lambda_i| < \lambda_1$, $i = 2, 3, \ldots, n$.

The assumption in (iii) is a reasonable one since it requires that there are two successive fertile age groups, which will usually be the case in realistic population models when the duration of each age group is sufficiently small.

■ ***EXAMPLE 1.26***

The Leslie matrix in (1.99) has $a_2 > 0$, $a_3 > 0$ so that the condition in (iii) is satisfied. The eigenvalues of L are $\lambda_1 = \frac{3}{2}$, $\lambda_2 = -1$, $\lambda_3 = -\frac{1}{2}$, showing that λ_1 is the strictly dominant eigenvalue with $|\lambda_1| > |\lambda_2|$, $|\lambda_1| > |\lambda_3|$.

We can now apply the properties (i)–(iii) to investigate what happens to the solution (1.93) when k becomes large. Suppose that L has been diagonalized as in (1.70), where

$$T^{-1}LT = D \tag{1.101}$$

and D is the diagonal matrix of eigenvalues of L. Multiply (1.101) on the left by T and on the right by T^{-1}, and remember that $TT^{-1} = T^{-1}T = I$, to get

$$L = TDT^{-1}$$

Hence

$$\begin{aligned} L^2 &= TDT^{-1}TDT^{-1} \\ &= TDIDT^{-1} \\ &= TD^2T^{-1} \end{aligned}$$

and repeating this process gives

$$L^k = TD^kT^{-1}, \quad k = 2, 3, 4, \ldots$$

Since D is a diagonal matrix with $\lambda_1, \lambda_2, \ldots, \lambda_n$ along the principal diagonal we have seen that

$$D^k = \begin{bmatrix} \lambda_1^k & 0 & 0 & \ldots & 0 \\ 0 & \lambda_2^k & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \lambda_n^k \end{bmatrix}, \quad k \geq 1$$

The solution in (1.93) therefore becomes

$$\begin{aligned} X(k) &= L^kX(0) \\ &= TD^kT^{-1}X(0) \end{aligned}$$

where dividing both sides by λ_1^k produces

$$\frac{1}{\lambda_1^k}X(k) = T\begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & (\lambda_2/\lambda_1)^k & 0 & \ldots & 0 \\ 0 & 0 & (\lambda_3/\lambda_1)^k & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & (\lambda_n/\lambda_1)^k \end{bmatrix} T^{-1}X(0)$$

Assuming λ_1 is the strictly dominant eigenvalue of L we have $|\lambda_i/\lambda_1| < 1$, so as in Section 1.2 it follows that

$$\left(\frac{\lambda_i}{\lambda_1}\right)^k \to 0 \text{ as } k \to \infty, \quad \text{for } i = 2, 3, \ldots, n$$

We have therefore shown that as $k \to \infty$

$$\frac{1}{\lambda_1^k} X(k) \to T \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 \end{bmatrix} T^{-1} X(0) \tag{1.102}$$

Recall that T is a constant matrix whose columns are eigenvectors of L corresponding to $\lambda_1, \ldots, \lambda_n$, so in particular the first column of T is the eigenvector u in (1.100). Let the first element of the vector $T^{-1}X(0)$ be denoted by c, a constant. Then the product on the right-hand side of (1.102) reduces to cu. To see how this works out, let's just look at the case $n = 3$. The right-hand side in (1.102) is

$$T \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} c \\ \bullet \\ \bullet \end{bmatrix} = T \begin{bmatrix} c \\ 0 \\ 0 \end{bmatrix}$$

where the dots indicate elements whose values don't matter. If the second and third columns in T are u_2 and u_3 this product is

$$[u, u_2, u_3] \begin{bmatrix} c \\ 0 \\ 0 \end{bmatrix} = cu$$

as required. We have therefore ended up with the result that as $k \to \infty$

$$\frac{1}{\lambda_1^k} X(k) \to cu$$

We can write this as

$$X(k) \to c\lambda_1^k u \tag{1.103}$$

which means that for large values of k, $X(k)$ behaves like $c\lambda_1^k u$ where c is a constant depending upon the initial vector $X(0)$.

If $\lambda_1 > 1$ equation (1.103) shows that the population is eventually increasing; if $\lambda_1 < 1$ the population is eventually decreasing (and so ends up at zero); and if $\lambda_1 = 1$ the population eventually stabilizes with zero growth. In this latter case $X(k) \to cu$, showing that for sufficiently large k the age distribution becomes a multiple of the eigenvector u corresponding to the eigenvalue $\lambda = 1$.

Replacing k by $k-1$ in (1.103) gives

$$X(k-1) \to c\lambda_1^{k-1} u$$

for large values of k, so we can conclude by comparing with (1.103) that

$$X(k) \to \lambda_1 X(k-1) \tag{1.104}$$

This means that after a sufficiently long time the age distribution vector $X(k)$ is λ_1 times the value $X(k-1)$ for the preceding time period, so that the *proportion* of females in each of the age groups becomes constant.

■ *EXAMPLE 1.27*

Return again to the Leslie matrix (1.99) whose strictly dominant eigenvalue was noted in Example 1.26 to be $\lambda_1 = \frac{3}{2}$. The eigenvector u in (1.98) is

$$u = \begin{bmatrix} 1 \\ \frac{1}{6} \\ \frac{1}{18} \end{bmatrix}$$

For large values of k we have seen in (1.104) that

$$X(k) \to \tfrac{3}{2} X(k-1)$$

so after a sufficiently long time ($k = N$, say) we can say that (to a close approximation)

$$X(N) = \tfrac{3}{2} X(N-1)$$

and similarly

$$X(N+1) = \tfrac{3}{2} X(N), \quad X(N+2) = \tfrac{3}{2} X(N+1), \quad \ldots$$

This means that the population *growth* becomes *constant*, since during *each* time period the number of females increases by 50%.

From (1.103) we see that after a long time has elapsed

$$X(k) \to c\left(\tfrac{3}{2}\right)^k \begin{bmatrix} 1 \\ \frac{1}{6} \\ \frac{1}{18} \end{bmatrix} \tag{1.105}$$

which shows that the numbers of females will be distributed among the three age groups in the ratios $1 : \frac{1}{6} : \frac{1}{18}$ or $18 : 3 : 1$. This converts into percentages within each of the three age groups of

$$\frac{18}{22} \times 100 = 81.8\%, \quad \frac{3}{22} \times 100 = 13.6\%, \quad \frac{1}{22} \times 100 = 4.6\%$$

EXERCISE 1.45 A population model in the form (1.92) has Leslie matrix

$$L = \begin{bmatrix} 0 & 13 & 12 \\ \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{2} & 0 \end{bmatrix}$$

Determine the eigenvalues of L using (1.95), and hence show that the ultimate tendency of the population is to double in size each time period. Show also that the distribution in the three age groups will approximate to 86.5%, 10.8%, 2.7% after a sufficiently long time has elapsed.

EXERCISE 1.46 A certain species of insect obeys the following rules for breeding and survival of females:

(i) $\frac{1}{16}$ survive their first birthday and live into a second year;
(ii) $\frac{1}{4}$ of these survive their second birthday and live into their third year;
(iii) by the end of the third year all the original insects are dead;
(iv) no insects are born until a female survives into its second year, when an average of seven new insects are produced, and this average drops to six in the third year of a female's life.

Obtain the model in the form (1.92), and deduce that the insects are doomed to extinction.

It's interesting to end our discussion of population models by giving an explicit expression for A^k when the $n \times n$ matrix A has distinct eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$.

It can be shown that

$$A^k = \lambda_1^k Z_1 + \lambda_2^k Z_2 + \cdots + \lambda_n^k Z_n, \quad k = 1, 2, 3, \ldots \tag{1.106}$$

where each Z_i is a constant $n \times n$ matrix. For example, when $n = 2$ then

$$Z_1 = \frac{A - \lambda_2 I}{\lambda_1 - \lambda_2}, \quad Z_2 = \frac{A - \lambda_1 I}{\lambda_2 - \lambda_1}$$

and when $n = 3$

$$Z_1 = \frac{(A - \lambda_2 I)(A - \lambda_3 I)}{(\lambda_1 - \lambda_2)(\lambda_1 - \lambda_3)}, \quad Z_2 = \frac{(A - \lambda_1 I)(A - \lambda_3 I)}{(\lambda_2 - \lambda_1)(\lambda_2 - \lambda_3)}, \quad Z_3 = \frac{(A - \lambda_1 I)(A - \lambda_2 I)}{(\lambda_3 - \lambda_1)(\lambda_3 - \lambda_2)} \tag{1.107}$$

You might be able to spot the pattern for a general value of n: Z_i is a product of matrices

$$(A - \lambda_1 I)(A - \lambda_2 I) \ldots (A - \lambda_n I)$$

excluding the factor $(A - \lambda_i I)$, divided by the product

$$(\lambda_i - \lambda_1)(\lambda_i - \lambda_2) \ldots (\lambda_i - \lambda_n)$$

excluding the factor $(\lambda_i - \lambda_i)$.

■ *EXAMPLE 1.28*

Return again to the 3×3 Leslie matrix L in (1.99) which was found to have eigenvalues $\lambda_1 = \frac{3}{2}$, $\lambda_2 = -1$, $\lambda_3 = -\frac{1}{2}$. We looked at the behaviour of L^k for large

values of k in Example 1.27, but we can now obtain an explicit expression for L^k for *any* value of the positive integer k. Using (1.107) with A equal to the matrix L in (1.99) and the stated values for the eigenvalues we have

$$Z_1 = \frac{(L+I)(L+\frac{1}{2}I)}{(\frac{3}{2}+1)(\frac{3}{2}+\frac{1}{2})}$$

$$= \frac{1}{5}\begin{bmatrix} 1 & 7 & 6 \\ \frac{1}{4} & 1 & 0 \\ 0 & \frac{1}{2} & 1 \end{bmatrix}\begin{bmatrix} \frac{1}{2} & 7 & 6 \\ \frac{1}{4} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

$$= \frac{1}{5}\begin{bmatrix} \frac{9}{4} & \frac{27}{2} & 9 \\ \frac{3}{8} & \frac{9}{4} & \frac{3}{2} \\ \frac{1}{8} & \frac{3}{4} & \frac{1}{2} \end{bmatrix} \tag{1.108}$$

EXERCISE 1.47 Verify that

$$Z_2 = \frac{4}{5}\begin{bmatrix} 1 & -4 & -6 \\ -\frac{1}{4} & 1 & \frac{3}{2} \\ \frac{1}{8} & -\frac{1}{2} & -\frac{3}{4} \end{bmatrix}, \quad Z_3 = \frac{4}{5}\begin{bmatrix} -\frac{1}{4} & \frac{1}{2} & 3 \\ \frac{1}{8} & -\frac{1}{4} & -\frac{3}{2} \\ -\frac{1}{8} & \frac{1}{4} & \frac{3}{2} \end{bmatrix}$$

From (1.105) we can now write

$$L^k = (\tfrac{3}{2})^k Z_1 + (-1)^k Z_2 + (-\tfrac{1}{2})^k Z_3 \tag{1.109}$$

The solution of the population equation is

$$X(k) = L^k X(0)$$

so that for a given initial state $X(0)$ we can determine $X(k)$. Furthermore, for large values of k (1.109) shows that

$$X(k) \to (\tfrac{3}{2})^k Z_1 X(0) \tag{1.110}$$

which is an explicit expression, unlike (1.105) which contains an unknown constant c.

EXERCISE 1.48 If

$$X(0) = \begin{bmatrix} 200 \\ 160 \\ 80 \end{bmatrix}$$

compute the products $Z_1X(0)$, $Z_2X(0)$, $Z_3X(0)$ using the expression (1.108) and Exercise 1.47. Hence determine $X(10)$ and the expression in (1.110). Compare the latter with (1.105).

EXERCISE 1.49 Obtain an expression for A^{100} using (1.105) when A is the matrix in Exercise 1.39.

EXERCISE 1.50 Use (1.105) to determine A^k when

$$A = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & 0 \end{bmatrix}$$

Hence show that the solution of

$$X(k+1) = AX(k), \quad X(0) = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$

approaches $\begin{bmatrix} \frac{7}{3} \\ \frac{7}{3} \end{bmatrix}$ as k becomes large.

EXERCISE 1.51 A certain population of insects is described by the standard model (1.92) with

$$L_1 = \begin{bmatrix} 0 & 2 & 3 \\ \frac{1}{2} & 0 & 0 \\ 0 & \frac{2}{3} & 0 \end{bmatrix}$$

In a different environment it is found that this changes to

$$L_2 = \begin{bmatrix} 0 & 3 & 4 \\ \frac{1}{3} & 0 & 0 \\ 0 & \frac{3}{4} & 0 \end{bmatrix}$$

Show that L_1 and L_2 have the same strictly dominant eigenvalue. Use your calculator to estimate this eigenvalue, and hence deduce that in either environment the population increases annually by about 32.5% after a sufficiently long time has elapsed.

In each case use (1.98) and (1.103) to obtain the ratios of the numbers in the three age groups after a long time has elapsed.

PROBLEMS

1.1 A sum of £200 is placed in a savings account which pays interest of $r\%$ compounded annually. If at the end of 5 years the account contains £270.23, determine r.

1.2 Consider again the aquarium model discussed in Examples 1.3, 1.6 and Exercise 1.14. Suppose that at the end of each week p units of water are removed from the aquarium, and $p+1$ units of fresh water are added so as to bring the water level back to normal (recall that 1 unit evaporates each week). Obtain the difference equation corresponding to (1.6). Deduce that after a long period of time the salt concentration approaches $(p+1)/p$ times the original concentration (it is assumed as before that n, the total number of units of water in the aquarium, is large). Notice that this agrees with the case $p = 2$ in Example 1.6.

1.3 Verify by direct substitution that

$$x_k = \alpha^k x_0 + \alpha^{k-1} + 2^2\alpha^{k-2} + 3^2\alpha^{k-3} + \cdots + k^2$$

is the general solution of

$$x_k = \alpha x_{k-1} + k^2, \quad k = 1, 2, 3, \ldots$$

What does the solution become when $\alpha = 1$?

1.4 The following model has been suggested to describe pollution of Lakes Erie and Ontario in North America:

(i) all the outflow from Lake Erie flows into Lake Ontario;
(ii) each year 38% of the water in Lake Erie and 13% of the water in Lake Ontario is replaced.

Let x_k and y_k denote the total amounts of pollution present in Lakes Erie and Ontario respectively at the start of year k.

For the period under consideration (i.e. from $k = 0$) new laws are introduced to protect the environment and these ensure that there is no further pollution of the lakes. Obtain the difference equations expressing x_{k+1} and y_{k+1} in terms of x_k and y_k. Solve these equations and show that the amount of pollution in Lake Erie is reduced to 10% of its original value after 5 years. Assuming that $x_0 = 3y_0$, show that to achieve the same reduction for Lake Ontario takes approximately 29 years.

1.5 Denote the Fibonacci numbers defined in Section 1.1 by $f_0 = 1$, $f_1 = 1$, $f_2 = 2$, and so on. Prove the following identities by the method of induction (see the appendix to this chapter):

(a) $f_0^2 + f_1^2 + f_2^2 + \cdots + f_n^2 = f_n f_{n+1}$, $n \geqslant 0$

(b) $f_1 + f_3 + f_5 + \cdots + f_{2n-1} = f_{2n} - 1$, $n \geqslant 1$

(c) $f_{2n+1}^2 = f_{2n} f_{2n+2} - 1$, $n \geqslant 0$.

1.6 Fibonacci numbers arise in connection with the following model of the behaviour of atoms of hydrogen gas.

A single electron belonging to an atom is initially in the ground level of energy (state 0) and is assumed to gain and lose energy alternately in succession. The rules are:

(i) when the gas gains radiant energy, all the electrons in state 1 rise to state 2; half those in state 0 rise to state 1 and half to state 2;
(ii) when the gas loses energy, all the electrons in state 1 fall to state 0; half those in state 2 fall to state 1 and half to state 0.

The histories of the states occupied by an electron are then as follows:

After an initial energy gain there are *two* possible histories, either 01 or 02.
There is then an energy loss, giving *three* histories: 010, 021, 020.
After the next energy gain there are *five* possible histories: 0101, 0102, 0212, 0201, 0202.

The sequence of events can be represented diagrammatically as in Figure 1.11.

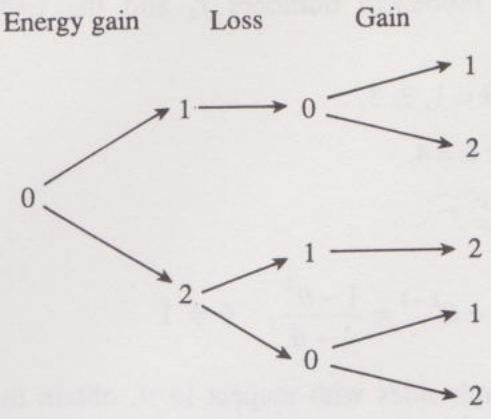


Figure 1.11

Extend this diagram for the next sequence of energy loss and gain, and verify that the numbers of different histories of the electron are the next two Fibonacci numbers.

1.7 The *binomial expansion formula* is

$$(1+a)^n = 1 + \binom{n}{1}a + \binom{n}{2}a^2 + \cdots + \binom{n}{n-1}a^{n-1} + a^n$$

where n is a positive integer and

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}, \quad r < n; \quad \binom{n}{n} = 1$$

$$\binom{n}{r} = 0, \quad r > n; \quad r! = 1 \times 2 \times 3 \times \cdots \times (r-1) \times r$$

Apply this to the expression (1.8) for the kth Fibonacci number x_k. Hence show that

$$x_k = \frac{1}{2^k}\left[\binom{k+1}{1} + \binom{k+1}{3}5 + \binom{k+1}{5}5^2 + \binom{k+1}{7}5^3 + \cdots\right]$$

Verify this is correct for $k = 1, 2, 3, 4, 5$.

1.8 The *Lucas numbers*

1, 3, 4, 7, 11, 18, 29, ...

also turn up in applications related to the natural world. They are defined by the same difference equation as for Fibonacci numbers, namely

$$L_{k+2} = L_{k+1} + L_k, \quad k = 0, 1, 2, \ldots$$

but with different initial values $L_0 = 1$, $L_1 = 3$. Use the solution obtained in (1.30) to show that

$$L_k = \left(\frac{1+\sqrt{5}}{2}\right)^{k+1} + \left(\frac{1-\sqrt{5}}{2}\right)^{k+1}, \quad k = 0, 1, 2, 3, \ldots$$

Check this result for $k = 4$.

1.9 Using the result in the previous problem, together with the expression in (1.8) for f_k, prove that the Fibonacci numbers f_k and the Lucas numbers L_k satisfy the relationship

$$f_{2k+1} = f_k L_k, \quad k = 1, 2, 3, \ldots$$

Check this result for $k = 4$.

1.10 Verify the identity

$$1 + \theta + \theta^2 + \cdots + \theta^{k-1} = \frac{1 - \theta^k}{1 - \theta}, \quad \theta \neq 1$$

By differentiating both sides with respect to θ, obtain the identity given in Exercise 1.15.

1.11 Wild plants propagate by self-seeding. Field observations of a certain species suggest the following rules:

(i) the plants flower and produce seeds either 1 year or 2 years after germination;
(ii) plants die after flowering;
(iii) 20% of the seeds produce plants that flower after 1 year, whereas 50% of the seeds produce plants that flower after 2 years, and the remainder (30%) of the seeds fail to produce plants that survive to produce more seeds;
(iv) on average each plant flowering after 1 year produces 350 seeds, compared with a figure 750 for plants flowering after 2 years.

Let s_k denote the number of seeds produced by flowers in year k. Show that

$$s_{k+2} = 70 s_{k+1} + 375 s_k, \quad k = 0, 1, 2, \ldots$$

If initially there are 10 seeds (i.e. $s_0 = 10$) deduce that $s_1 = 700$ and obtain a general expression for s_k.

1.12 Solve the equation

$$x_k = x_{k-1} + k^2, \quad k = 1, 2, 3, \ldots$$

subject to $x_0 = 0$ by substituting

$$x_k = ak^3 + bk^2 + ck$$

Equate coefficients of powers of k to obtain three equations for the three constants a, b, c and hence show that $a = \frac{1}{3}$, $b = \frac{1}{2}$, $c = \frac{1}{6}$.

Notice that

$$x_1 = 1^2, \quad x_2 = x_1 + 2^2 = 1^2 + 2^2$$
$$x_3 = x_2 + 3^2 = 1^2 + 2^2 + 3^2, \ldots$$

so it follows from the expression for x_k that

$$1^2 + 2^2 + 3^2 + \cdots + k^2 = \tfrac{1}{3}k^3 + \tfrac{1}{2}k^2 + \tfrac{1}{6}k$$
$$= \tfrac{1}{6}k(k+1)(2k+1)$$

Use this method to solve

$$x_k = x_{k-1} + (2k-1)^2, \quad k = 1, 2, 3, \ldots$$

subject to $x_0 = 0$. Hence prove that

$$1^2 + 3^2 + 5^2 + \cdots + (2k-1)^2 = \frac{k(4k^2-1)}{3}$$

1.13 Modify the method used in the preceding problem to solve

$$x_k = x_{k-1} + k^3, \quad k = 1, 2, 3, \ldots$$

subject to $x_0 = 0$. Hence obtain an expression for the sum of the cubes of the integers from 1 to k.

1.14 A game with two players (A and B) involves tossing coins which have an equal probability of coming up heads or tails. The rules are:

(i) if a coin comes up heads, A gives B one coin
(ii) if a coin comes up tails, B gives A one coin.

The winner is the player who first ends up with all the coins.

Suppose that initially A has k coins and B has $N-k$ coins. Let p_k be the probability that A wins, so that $p_0 = 0$ and $p_N = 1$. Show that

$$2p_{k+1} = p_{k+2} + p_k, \quad k = 0, 1, 2, \ldots$$

Solve this equation to obtain $p_k = k/N$.

1.15 A coin has a probability p of coming up heads and $q = 1 - p$ of coming up tails when tossed. A gambler wins £1 if the coin shows heads, and loses £1 if it shows tails. The gambler begins with £a and aims to quit when he or she has £b (with $b > a$). If the gambler loses all his or her money before achieving this goal then the gambler is said to be *ruined*.

Let p_k denote the probability of eventual ruin when the gambler has £k. Notice that $p_0 = 1$ (the gambler is already ruined) and $p_b = 0$ (the gambler has won). Show that

$$pp_{k+1} - p_k + (1-p)p_{k-1} = 0, \quad k = 1, 2, 3, \ldots$$

Solve this difference equation for each of the cases $p \neq q$ and $p = q$. Hence show that the probability of eventual ruin with the original stake of £a is

$$p_a = \frac{1}{1-\alpha^b}(\alpha^a - \alpha^b), \quad \alpha = \frac{q}{p} \neq 1$$

$$= 1 - \frac{a}{b}, \quad p = q$$

Deduce that when the coin is unbiased (i.e. $p = q$) the gambler has a 50% chance of being ruined when trying to double the original stake, and a 75% chance when trying to quadruple this stake.

1.16 Owing to restricted water supplies a farmer can irrigate his or her fields only from 9 p.m. to 9 a.m. During this period the farmer adds a quantity c of water to the topsoil. However, during the period from 9 a.m. to 9 p.m. half the total water in the topsoil is lost through evaporation and absorption.

Let x_k denote the amount of water in the topsoil at the *end* of the kth 12 hour period, starting with x_0 at 9 p.m. on the first day. Show that

$$x_{k+2} = \tfrac{1}{2}x_k + \tfrac{1}{4}c[3-(-1)^k], \quad k = 0, 1, 2, \ldots$$

by considering separately the cases k odd and k even.

Solve this equation, and hence deduce that when this irrigation programme has been followed for a long time, x_k essentially oscillates between c and $2c$.

1.17 Use the z-transform to obtain the solution of

$$x_{k+1} = \alpha x_k + c^k, \quad k = 0, 1, 2, \ldots$$

for each of the cases $c \neq \alpha$, $c = \alpha$.

1.18 Return to the simple model of a national economy described in Exercise 1.10. If $\alpha = \frac{1}{2}$, $\beta = 1$, government spending is at a constant level $G_k = G$ (all k) and $I_0 = 2G$, $I_1 = 3G$, show by using the z-transform that

$$I_k = 2\left[1 + \left(\frac{1}{\sqrt{2}}\right)^k \sin\left(\tfrac{1}{4}k\pi\right)\right]G, \quad k = 0, 1, 2, \ldots$$

You will need the fact that

$$Z\left[\left(\frac{1}{\sqrt{2}}\right)^k \sin\left(\tfrac{1}{4}k\pi\right)\right] = \frac{z}{2z^2 - 2z + 1}$$

The assumptions of the model therefore produce a national income which oscillates because of the sine term, and as $k \to \infty$ the national income approaches twice government expenditure.

1.19 (a) The *hyperbolic functions* are defined by

$$\cosh x = (e^x + e^{-x})/2, \quad \sinh x = (e^x - e^{-x})/2$$

Show that $\cosh^2 x - \sinh^2 x = 1$. If w is the positive solution of the equation $\cosh x = 3/2$, show that $\sinh w = \sqrt{5}/2$.

Given the transforms

$$Z(\sinh wk) = \frac{z \sinh w}{z^2 - 2z\cosh w + 1}$$

$$Z(\cosh wk) = \frac{z(z - \cosh w)}{z^2 - 2z\cosh w + 1}$$

show that

$$Z(\cosh wk + c \sinh wk) = \frac{z^2 - z(3/2 - c\sqrt{5}/2)}{z^2 - 3z + 1}$$

where c is a constant.

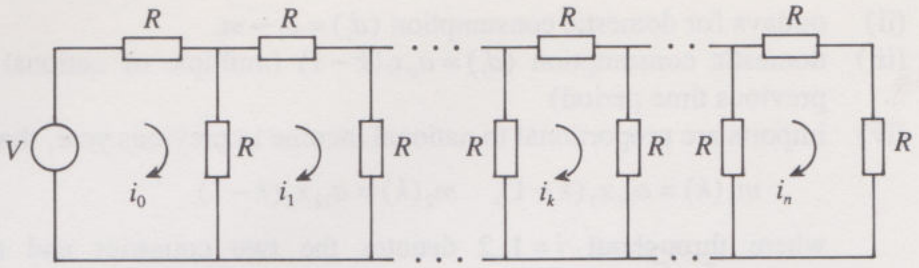


Figure 1.12

(b) A so-called ladder network of resistors R is shown in Figure 1.12. The applied voltage is V and i_k is the current in the kth loop. It can be shown by what is known as Kirchhoff's law that

$$i_{k+2} - 3i_{k+1} + i_k = 0, \quad k = 0, 1, 2, \ldots$$

with $i_1 = 2i_0 - V/R$. Obtain an expression for the z-transform of i_k. Use the result in part (a) to determine i_k in terms of i_0, w, V and R.

1.20 The definition (1.41) of the z-transform of a sequence x_k, $k = 0, 1, 2, \ldots$, states that

$$Z(x_k) = x_0 + \frac{x_1}{z} + \frac{x_2}{z^2} + \frac{x_3}{z^3} + \cdots$$

Differentiate both sides with respect to z and hence deduce that

$$Z(kx_k) = -z \frac{\mathrm{d}}{\mathrm{d}z} Z(x_k)$$

Use this result to show that

$$Z(kc^k) = \frac{cz}{(z-c)^2}$$

given that

$$Z(c^k) = \frac{z}{z-c}$$

where c is a constant.

1.21 A simple economic model of 'supply and demand' assumes that if the supply of a commodity in year k is s_k then the price is $p_k = a - bs_k$, where a and b are positive constants (this implies that the price declines if the supply increases). It is also assumed that the supply in year $k+1$ is proportional to the price in the previous year, that is $s_{k+1} = cp_k$ where c is a positive constant.

Obtain the difference equation satisfied by p_k and find its general solution. Deduce that if $bc < 1$ then the price stabilizes after a sufficiently long time has elapsed.

1.22 Consider the following simple model of trade between two countries only, where the exports of one are the imports of the other. The assumptions are:

(i) national income (x_i) = consumption outlays (c_i) + net investment (v_i) + exports (e_i) − imports (m_i)

(ii) outlays for domestic consumption $(d_i) = c_i - m_i$
(iii) domestic consumption $(d_i) = a_{ii}x_i(k-1)$ (multiple of national income in previous time period)
(iv) imports are proportional to national income in previous year, that is

$$m_1(k) = a_{21}x_1(k-1), \quad m_2(k) = a_{12}x_2(k-1)$$

where throughout $i = 1, 2$ denotes the two countries and the a_{ij} are constants.

Show that

$$\begin{bmatrix} x_1(k) \\ x_2(k) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1(k-1) \\ x_2(k-1) \end{bmatrix} + \begin{bmatrix} v_1(k) \\ v_2(k) \end{bmatrix}, \quad k = 0, 1, 2, \ldots$$

1.23 Consider a simplified cattle ranching model where the numbers of females in year k are

$x_1(k)$ = number of 1-year-olds ('young')
$x_2(k)$ = number of 2-year-olds ('mature')
$x_3(k)$ = number of 3-year-olds and older ('old')

In the absence of slaughtering the assumptions concerning breeding and mortality are as follows:

(i) young females do not breed;
(ii) a mature female produces on average 0.8 young cattle per year;
(iii) an old female produces on average 0.4 young cattle per year;
(iv) only old cattle die, at the rate of 30% per year.

Show that

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \\ x_3(k+1) \end{bmatrix} = \begin{bmatrix} 0 & 0.8 & 0.4 \\ 1 & 0 & 0 \\ 0 & 1 & 0.7 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \\ x_3(k) \end{bmatrix}, \quad k = 0, 1, 2, \ldots$$

Notice that in this case the matrix does not have the Leslie form (1.89), since we have not assumed that the maximum lifespan of cattle is 3 years.

1.24 In the car rental model described in Example 1.23, after several years have passed it is found that 80% of cars rented at B are dropped off at L, whereas only 25% of those rented at L are dropped off at B. Otherwise, as before, cars are returned to the originating office. Find the new equilibrium situation in this case, assuming that the total company fleet has grown to 210 cars.

1.25 Two TV channels show competing newscasts in the same time slot every evening. Audience research amongst those who always watch the news shows that if a viewer watches Channel X on one evening there is a 50% chance of switching to Channel Y the following evening. However, the programme on Channel Y is more enjoyable, so there is only a 40% chance of returning to Channel X. Let x_k, y_k be the probabilities that a viewer is watching Channel X or Y respectively on the kth day. Show that

$$x_{k+1} = \tfrac{1}{2}x_k + \tfrac{2}{5}y_k, \quad y_{k+1} = \tfrac{1}{2}x_k + \tfrac{3}{5}y_k, \quad k = 0, 1, 2, \ldots$$

Write these equations in matrix form and obtain the general solution. Deduce that after a long enough time has elapsed the probabilities that a viewer watches the news on Channel X or Channel Y are $\frac{4}{9}$ and $\frac{5}{9}$ respectively.

1.26 The following model has been proposed for a trout fish farm. In the four stages in the life of a trout, define at year k:

$x_1(k)$ = number of eggs
$x_2(k)$ = number of fry
$x_3(k)$ = number of young
$x_4(k)$ = number of adults

Define also

$u_1(k)$ = number of eggs added artificially
$u_2(k)$ = number of young removed for stocking streams

From observations over several years it is found that:

(i)

$$\begin{pmatrix}\text{number of}\\ \text{eggs at}\\ \text{year } k\end{pmatrix} = \begin{pmatrix}\text{number of}\\ \text{eggs added}\\ \text{in year } k\end{pmatrix} + \begin{pmatrix}\text{number of}\\ \text{eggs laid by}\\ \text{adults}\\ \propto x_4(k)\end{pmatrix} - \begin{pmatrix}\text{number of}\\ \text{eggs eaten by}\\ \text{fry}\\ \propto x_2(k)\end{pmatrix} - \begin{pmatrix}\text{number of}\\ \text{eggs eaten by}\\ \text{young}\\ \propto x_3(k)\end{pmatrix}$$

(ii) a constant proportion of eggs in year k survive to become fry in the following year;

(iii) a constant proportion of fry in year k survive to become young in the following year;

(iv)

$$\begin{pmatrix}\text{number of adults}\\ \text{in year } k\end{pmatrix} \propto \begin{pmatrix}\text{number of young + number of adults}\\ \text{- number of young removed}\end{pmatrix}\Big\}\ \text{in previous year}$$

Show that

$$\begin{bmatrix}x_2(k+1)\\ x_3(k+1)\\ x_4(k+1)\end{bmatrix} = \begin{bmatrix}-a_1 & -a_2 & a_3\\ a_4 & 0 & 0\\ 0 & a_5 & a_5\end{bmatrix}\begin{bmatrix}x_2(k)\\ x_3(k)\\ x_4(k)\end{bmatrix} + \begin{bmatrix}a_6 & 0\\ 0 & 0\\ 0 & -a_5\end{bmatrix}\begin{bmatrix}u_1(k)\\ u_2(k)\end{bmatrix}, \quad k = 0, 1, 2, \ldots$$

where the a_i are positive constants.

1.27 For a certain population of wild animals, censuses taken twice yearly at the end of April and at the end of October reveal that no animal lives for more than 18 months. This period is divided into three time intervals each of length 6 months. Let $x_1(k)$, $x_2(k)$ and $x_3(k)$ be the numbers of animals in the kth 6 monthly period in each of the age groups 0–6 months (young), 6–12 months (mature) and 12–18 months (old) at

the time of the April census. Let $y_1(k)$, $y_2(k)$, $y_3(k)$ be the corresponding numbers at the time of the October census. Observations show that

$$\underbrace{\begin{bmatrix} y_1(k) \\ y_2(k) \\ y_3(k) \end{bmatrix}}_{Y(k)} = \begin{bmatrix} 0 & 4 & 6 \\ \frac{2}{3} & 0 & 0 \\ 0 & \frac{2}{3} & 0 \end{bmatrix} \underbrace{\begin{bmatrix} x_1(k) \\ x_2(k) \\ x_3(k) \end{bmatrix}}_{X(k)}$$

and also

$$X(k+1) = \begin{bmatrix} 0 & 1 & 3 \\ \frac{1}{2} & 0 & 0 \\ 0 & \frac{8}{9} & 0 \end{bmatrix} Y(k)$$

(a) Obtain the matrix difference equation

$$Y(k+1) = AY(k), \quad k = 0, 1, 2, \ldots$$

and determine the eigenvalues of A. Similarly obtain the expression

$$X(k+1) = BX(k)$$

and verify that B has the same eigenvalues as A.

(b) Consider the equilibrium situation which is attained after a long time has elapsed. Show that in this case the population increases annually by a factor $\frac{8}{3}$.

(c) Because of this population growth the animals have become a pest, so it is decided to operate an extermination programme. In this scheme $\frac{15}{16}$ of young animals, $\frac{1}{2}$ of mature animals and $\frac{23}{32}$ of old animals will be killed each October. Show that after a sufficiently long time the population will decrease by 50% every year (and so will eventually become extinct).

1.28 A simplified model for the wild buffalo population in the American west in 1830 has been suggested as follows. Let F_k, M_k be the numbers of adult female and male buffalo respectively at the start of year k, where $k = 0$ corresponds to 1830. On the basis of observations the following rules apply:

(i) 5% of adults die each year;
(ii) the animals reach maturity at age 2 years;
(iii) the number of new adult females alive at the beginning of year $k+2$, taking into account infant mortality, is 12% of F_k;
(iv) more male calves are born than female, the figure corresponding to (iii) being 14% of F_k.

Show that

$$F_{k+2} = 0.95F_{k+1} + 0.12F_k$$

$$M_{k+2} = 0.95M_{k+1} + 0.14F_k, \quad k = 0, 1, 2, \ldots$$

Use the z-transform to obtain an expression for F_k in terms of F_0 and F_1. Hence deduce that when k is sufficiently large the population increases by 6.3% per year.

1.29 Consider the Leslie matrix L in (1.89). Show that the average number of daughters born to a single female during her expected lifetime is

$$d = a_1 + a_2 b_1 + a_3 b_1 b_2 + \cdots + a_n b_1 b_2 \ldots b_{n-1}$$

Deduce that L has an eigenvalue equal to 1 if and only if $d = 1$.

1.30 A model of a population of mice uses the following assumptions:

(i) mice do not mate until they are 1 month old ('mature');
(ii) each pair of mature mice present at the end of 1 month produces two new pairs by the end of the next month;
(iii) no mice die.

Let x_k denote the number of pairs of mice at the end of the kth month. Show that

$$x_{k+2} - x_{k+1} - 2x_k = 0, \quad k = 0, 1, 2, \ldots$$

Obtain the solution of this equation subject to the conditions $x_0 = 2$, $x_1 = 4$. How long will it take for the mouse population to exceed 500 pairs?

1.31 A particle is moving along the x-axis in the direction of increasing x. Its x-coordinate after k seconds have elapsed since starting off is x_k. The distance the particle travels from time k to time $k+1$ is equal to twice the distance it travelled from time $k-1$ to time k. Show that

$$x_{k+2} - 3x_{k+1} + 2x_k = 0, \quad k = 0, 1, 2, \ldots$$

and find the expression for x_k subject to $x_0 = 1$, $x_1 = 5$.

1.32 An investor has £1000 and wishes to turn this into £1300 as quickly as possible by investing it in one of the following accounts:

(i) annual interest of 8%;
(ii) APR of $7\frac{1}{4}$ compounded quarterly;
(iii) APR of $6\frac{1}{2}$ compounded monthly.

Which option should be selected?

APPENDIX: PROOF BY INDUCTION

The method of induction is an important way of proving results which you have *guessed* may be correct. For example, let S_n be the sum of the first n positive integers. By direct addition

$$S_1 = 1, \quad S_2 = 1 + 2 = 3, \quad S_3 = 1 + 2 + 3 = 6, \quad S_4 = 1 + 2 + 3 + 4 = 10$$

and similarly $S_5 = 15$, $S_6 = 21$. You may be able to spot a pattern: what is happening is that

$$S_1 = \frac{1 \times 2}{2}, \quad S_2 = \frac{2 \times 3}{2}, \quad S_3 = \frac{3 \times 4}{2}, \quad S_4 = \frac{4 \times 5}{2}$$

and sure enough this also works for S_5 and S_6:

$$S_5 = \frac{5\times 6}{2}, \quad S_6 = \frac{6\times 7}{2}$$

It therefore *looks* as though in general

$$S_n = \frac{n(n+1)}{2} \tag{A1}$$

but it is important to realize that this remains a *guess* at this point. We can't be sure that the formula (A1) is correct even if we verify that it works for n going from 1 to 100, or even from 1 to 1000.

It is a *fallacy* to assume that if a formula works for several (or even very many) values of n then it must be true for *all* values of n. A simple illustration of this fallacy is provided by the formula

$$a_n = \tfrac{1}{24}(n^4 - 6n^3 + 23n^2 - 18n + 24) \tag{A2}$$

It is easy to check with a calculator that

$$a_1 = \tfrac{1}{24}(1 - 6 + 23 - 18 + 24) = 1$$
$$a_2 = \tfrac{1}{24}(16 - 6\times 8 + 23\times 4 - 18\times 2 + 24) = 2$$
$$a_3 = 4 = 2^2, \quad a_4 = 8 = 2^3, \quad a_5 = 16 = 2^4$$

and it seems 'obvious' that for any positive integer n

$$a_n = 2^{n-1} \tag{A3}$$

However, you can also check that substituting $n=6$ into (A2) produces $a_6 = 31$, which is *not* 2^5 $(= 32)$. Thus, although (A3) is correct for $n = 1, 2, 3, 4, 5$, it is *not* true for all values of n.

Let's go back to the formula (A1). We certainly know that it is correct for $n = 1$. What we need to do is to prove:

If the formula is true for n equal to any positive integer N, then it is also true for $n = N + 1$. (*)

This is quite easy: if we assume

$$S_N = \frac{N(N+1)}{2} \tag{A4}$$

then clearly from the definition of S_N we have

$$S_{N+1} = S_N + (N+1) \tag{A5}$$

$$= \frac{N(N+1)}{2} + (N+1)$$

$$= \frac{(N+1)(N+2)}{2} \tag{A6}$$

This result (A6) is *exactly* the same as what we get in (A4) if we replace N by $N+1$. We have therefore established the condition (*) in this example: if (A1) holds for $n=N$ then it is also true for $n=N+1$. However, we know that (A1) is correct for $n=1$, so it must also be true for $n=2$; and since it is true for $n=2$, the condition (*) tells us it must be true for $n=3$. You should now realize that we can proceed in this way *for ever*: the correctness of (A1) for $n=3$ implies it is true for $n=4$, and similarly for $n=5$, and so on. Hence we can conclude that (A1) is indeed true for *any* value of the positive integer n.

In general, the *method of induction* to prove that a guessed formula S_n is correct consists of showing that it is correct for some particular value $n=a$ (in the example above $a=1$) and then establishing that (*) holds, from which it follows that S_n holds for all integers $n \geqslant a$.

If you are still a bit mystified, let's see what would happen if we guessed the *wrong* formula for the sum S_n of the first n positive integers. Suppose we thought that

$$S_n = n^2 - n + 1 \tag{A7}$$

This certainly works for $n=1$ and $n=2$, since

$$S_1 = 1^2 - 1 + 1 = 1, \quad S_2 = 2^2 - 2 + 1 = 3$$

If we assume this is true for $n=N$ then we have

$$S_N = N^2 - N + 1 \tag{A8}$$

and hence

$$\begin{aligned} S_{N+1} &= S_N + (N+1) \\ &= N^2 - N + 1 + N + 1 \\ &= N^2 + 2 \end{aligned} \tag{A9}$$

However, if we replace N by $N+1$ in (A8) we get

$$\begin{aligned} S_N &= (N+1)^2 - (N+1) + 1 \\ &= N^2 + 2N + 1 - N \\ &= N^2 + N + 1 \end{aligned}$$

which is *not* the same as (A9). This shows that the statement (*) does *not* hold, so if (A7) holds for $n=2$ it does *not* follow that it is also true for $n=3$ — indeed (A7) gives $S_3 = 3^2 - 3 + 1 = 7$ instead of the correct value $S_3 = 6$; the fact that (A7) works for both $n=1$ and $n=2$ is just a fluke.

As another example, consider the formula

$$S_n = 2^{2n} - 1 \tag{A10}$$

which gives

$$S_1 = 2^2 - 1 = 3, \quad S_2 = 2^4 - 1 = 15, \quad S_3 = 2^6 - 1 = 63$$

and if you compute a few more values of S_n you will find that each is divisible by 3.

To prove by induction that this is true for all values of the positive integer n, we need to show that (*) holds. That is, if

$$S_N = 2^{2N} - 1$$

is divisible by 3, we must show that S_{N+1} is also divisible by 3. Because S_N is divisible by 3, we can write it in the form $3k$ where k is a positive integer, so that

$$2^{2N} - 1 = 3k$$

or $2^{2N} = 3k + 1$. Next, setting $n = N + 1$ in (A10) gives

$$\begin{aligned} S_{N+1} &= 2^{2(N+1)} - 1 \\ &= 2^{2N+2} - 1 \\ &= 2^2.2^{2N} - 1 \\ &= 4.2^{2N} - 1 \\ &= 4(3k+1) - 1 = 12k + 3 \end{aligned}$$

which is again divisible by 3. We have therefore established that (*) does indeed hold, and since S_1 is divisible by 3 it follows that S_n is divisible by 3 for *all* values of the positive integer n.

You will need the technique described in this appendix to solve Exercise 1.8 and Problem 1.5 involving some properties of Fibonacci numbers.

EXERCISE A1 Prove by the method of induction that each of the following results holds for all values of the positive integer n:

(a) $1^2 + 2^2 + 3^2 + 4^2 + \cdots n^2 = n(n+1)(2n+1)/6$ (see also Problem 1.12).
(b) $2^{3n} - 1$ is divisible by 7.

EXERCISE A2 Let S_n be the sum of the odd integers from 1 to $2n - 1$. By evaluating S_1, S_2, S_3, S_4 guess a formula for S_n. Use the method of induction to prove that your result is true for all values of n.

FURTHER READING

BARNETT, S. 1990. *Matrices: Methods and Applications*. Oxford University Press, Oxford.

CADZOW, J.A. 1973. *Discrete-Time Systems*. Prentice Hall, Englewood Cliffs, NJ.

DIERKER, P.F. and VOXMAN, W.L. 1986. *Discrete Mathematics*. Harcourt Brace Jovanovich, San Diego.

EPP, S.S. 1990. *Discrete Mathematics with Applications*. Wadsworth, Belmont, CA.

GABEL, R.A. and ROBERTS, R.A. 1973. *Signals and Linear Systems*. Wiley, New York.

GARLAND, T.H. 1987. *Fascinating Fibonaccis, Mystery and Magic in Numbers*. Dale Seymour, Palo Alto, CA.

HUNTLEY, H.E. 1970. *The Divine Proportion. A Study in Mathematical Beauty*. Dover, New York.

KELLEY, W.G. and PETERSON, A.C. 1991. *Difference Equations. An Introduction with Applications*. Academic Press, London.

LIGHTHILL, M.J. (Ed.). 1978. *Newer Uses of Mathematics*. Penguin, London, Chapter 2.

LINES, M.E. 1990. *Think of a Number*. Adam Hilger, Bristol, Chapter 2.

MAURER, S.B. and RALSTON, A. 1991. *Discrete Algorithmic Mathematics*. Addison-Wesley, Reading, MA.

RORRES, C. and ANTON, H. 1991. *Applications of Linear Algebra*, 6th Edition. Wiley, New York.

SANDEFUR, J.T. 1990. *Discrete Dynamical Systems. Theory and Applications*. Oxford University Press, Oxford.

2

Supermarket Barcodes, Pictures From Space, Compact Discs

2.1 INTRODUCTION AND EXAMPLES

Almost every product we buy in a supermarket has a barcode on it, like that shown in Figure 2.1.

The bars are designed to be read by a laser scanning system, producing the string of numbers which identifies the item (in this case, a tube of hair cream). Turn over any book, and in the bottom right-hand corner of the back cover you will find another barcode, like that in Figure 2.2.

This number is called the International Standard Book Number (ISBN), and every published book has its own unique number. Barcodes are very widely used in commerce and industry, and play an important role in the processing of management information.

The second part of this chapter's title refers to one of the most impressive technological feats of the twentieth century: the transmission of pictures to Earth by spacecraft visiting other planets in the solar system. The first, rather blurred, pictures of Mars were received in 1965, but the Mariner 9 spacecraft in 1971 sent television signals from Mars to produce pictures of excellent quality back on Earth. Yet the

Figure 2.1

Figure 2.2

power of the transmitter was so small (only 20 watts) that what happened was like being able to see a car stoplight on Mars, which is 135 million km away! Compare this with a commercial television transmitter which needs about 35 000 watts for a range of around 80 km. In 1976 colour photographs of Mars were obtained, and in the 1980s spectacular pictures of Saturn, Uranus and Neptune were produced. The basic principle used is to break down the pictures into a large number of very small elements, and transmit these as numerical data. The way in which such data are processed at the receiving end so as to get rid of distortions and interference forms the topic of this chapter.

Let's look at a simple problem in which representing information in numerical form can be useful.

■ *EXAMPLE 2.1*

Suppose we have a list of eight names in alphabetical order:

Ann, Bob, Carol, Dave, Ellen, Fred, Gill, Harry

I am thinking about one name on this list, and you have to determine which one it is by asking me questions, to which I will answer only 'yes' or 'no'. What is the smallest number of questions you need to ask in order to get the right name?

The answer is 'three': first ask me whether the name is in the first half of the list – this narrows it down to four names. Then ask me whether the name is in the first half of this shorter list – this narrows it down to two names. Finally, ask me whether the name is the first in this last pair. For example, if I am thinking of 'Carol' my first answer is 'yes' – the name is in the first four (Ann to

Dave); my second reply is 'no', revealing that the name is either the third or fourth (Carol or Dave); finally I say 'yes', so that you can identify the name as the third on the original list.

Suppose I now change the rules of the game. You ask questions as before, but instead of answering one at a time I only give you my answers after you have asked all your questions. Do you need to ask more than three questions? In fact, perhaps surprisingly, you still only need three questions to identify the correct name. The easiest way to see this is to first assign a number to each of the names in the list, going from 0 for Ann, 1 for Bob, up to 7 for Harry. The reason for going from 0 to 7 rather than from 1 to 8 is that we can represent each of the numbers in *binary form*. You are familiar with the way decimal numbers work: the number 751, for example, means 7 hundreds, 5 tens, and 1 unit, or $7 \times 10^2 + 5 \times 10^1 + 1 \times 10^0$. The position of the digit tells you what power of the *base* 10 it corresponds to. In exactly the same way, binary numbers use base 2, so there are only two digits 0 and 1, called *bits* (short for *binary digits*). For example, the binary number 10101 means

$$1 \times 2^4 + 0 \times 2^3 + 1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0$$

which is $16 + 4 + 1 = 21$ in decimal form. Most scientific pocket calculators have a button which converts binary numbers to decimal numbers, and vice versa.

Going back to our original problem, the binary representations of the decimal numbers 0 to 7 are as follows:

decimal	0	1	2	3	4	5	6	7
binary	000	001	010	011	100	101	110	111
name	Ann	Bob	Carol	Dave	Ellen	Fred	Gill	Harry

Notice that we only need 3 bits for each of the numbers. The binary form of 8 is 1000, so if we had numbered the names from 1 to 8 we would have had to use four digits for Harry.

To play the second version of the game, all you have to do is ask me if each digit is 0 reading from left to right. Remember my choice was Carol, third on the list, and having the code number 010, as shown in the table above. So my answer, after you have asked the three questions, would be 'yes, no, yes' – in fact, the same answers as in the original game.

We say that the information about the list of names has been *encoded* into binary form. The numbers 000, 001, 010 and so on are called *codewords*, and the set of all these is a *code*.

EXERCISE 2.1 Analyze the game described in Example 2.1, in the cases when there are 16 names, or when there are 32 names.

Human counting generally uses the decimal system with base 10, although base 12 occurs in the old Imperial measuring system of feet and inches, and in the old British coinage of pennies and shillings. The base 60 goes back to Babylonian times, and is still firmly with us for time and angle measurement (seconds, minutes, hours). However, computers are built from electronic components which can be in either one state or another (e.g. an on–off switch). It is therefore natural to use binary numbers

to represent the two states. The decimal system undoubtedly has its origins in the fact that humans have five digits (four fingers and a thumb) on each of their two hands; so perhaps we can imagine that a computer has only one finger on each hand!

The third component in the title of the chapter is the now-familiar *compact disc* (CD) for reproducing sound. The old-fashioned technology of the long-playing record relies on the sound being converted into a long wiggly groove which goes round and round the record. The wiggles are traced by a stylus, whose movements are turned back into music. Although ingenious, this technique unfortunately allows too much scope for distortion and loss of quality in the reproduction. In the technology of the compact disc, which was introduced in 1982, the musical sounds are decomposed into tiny individual parts which are converted into digital form: in just 1 second a CD player processes 1 460 000 bits of audio information. The bits are read off the disc by a laser beam. However, even with the most careful manufacturing and handling procedures, faults still occur on CDs. The reason why, despite such flaws, the music sounds so authentic and free from 'clicks' and other unwanted background noises is that the CD contains about twice as many more bits of non-audio information. This extra information is used to process the music on its way to your ears, so that it ends up sounding virtually perfect. In particular, some of these extra bits are used to correct errors, which is what this chapter is all about. The actual error-correcting scheme for CDs was invented at Philips Research Laboratories in Eindhoven in the late 1970s, and emphasizes that in this important technological development the electronic engineers could not have succeeded without the contribution of the mathematicians!

Incidentally, we shall not be discussing codes which are used by spies and others to maintain secrecy – this is another interesting branch of recent applied mathematics called *cryptography*.

In general we wish to send information which has been converted into numerical form along some communication channel as reliably as possible. This might be a telephone line, a satellite communication link or a magnetic disk used with a computer. Suppose we are using a binary representation of the data (i.e. a string of zeros and ones), so that what is happening can be as represented in Figure 2.3.

The message has some extra *check* bits appended onto it by the encoder device (in some cunning way) so as to produce a *codeword*. During transmission, so-called 'noise' (i.e. external interference) causes errors to occur. The sources of the noise could be, for example, electrical faults or disturbances, lightning, radiation or human error. The string of zeros and ones which is received is no longer the same as the original codeword – it might be a different codeword, or it might not be a codeword at all, in which case it is called simply a *word*. The fundamental idea of an error-correcting code is to utilize the extra bits added by the encoder in such a way that after decoding the received word the original message can be recovered. What we shall be looking at in this chapter is how to construct some of these cunning schemes for adding the check bits.

We are already familiar with a similar idea in written language. There is often sufficient redundancy already present in the structure of the language itself to enable

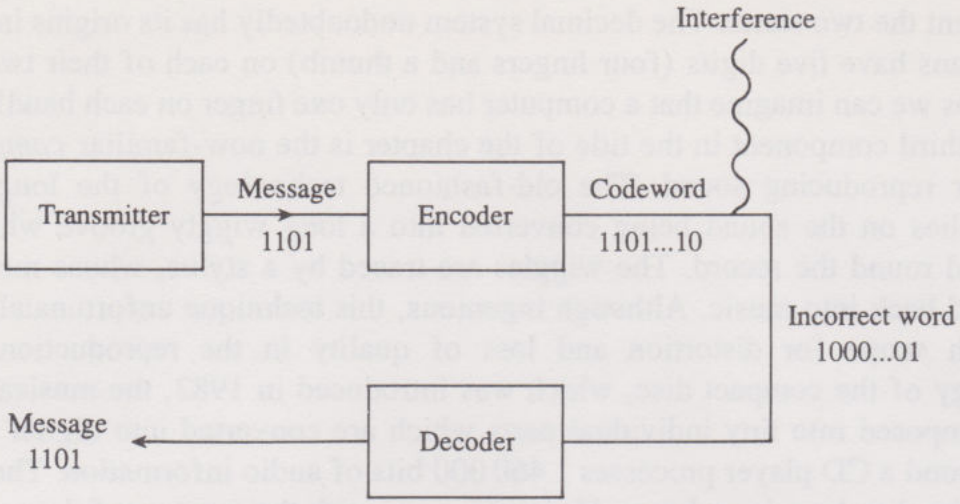


Figure 2.3

us to guess correctly what is meant even if there are spelling errors, or perhaps vowels omitted. For example, we can still understand 'Ystrdy ws cldy'; or we can still make sense of 'Tomorrow the weather will be five', since we can be pretty sure that there has been a typing error and 'five' should have been 'fine'. However, if a clothing warehouse received a telex message 'Supply 1000 shurts', there is no way of telling whether 'shirts' or 'shorts' are required. Even worse, if a bank clerk types in an incorrect account number during a financial transaction, then you could find money has been withdrawn from your account instead of somebody else's (Murphy's law suggests that it is unlikely that you would actually benefit from such a mistake!).

■ *EXAMPLE 2.2*

Suppose we want to send four commands to a robot arm involved in some manufacturing process: UP, DOWN, LEFT, RIGHT. We could use the following codewords:

UP	DOWN	LEFT	RIGHT
00	01	10	11

However, if even a single error occurs during transmission then there is no way of detecting that an error has occurred, as the received message is perfectly legitimate. For example, if 01 is sent, but is corrupted so that 11 is received, then the arm moves to the right instead of downwards.

When someone speaks to us and we don't quite catch what they say, then we ask them to repeat it. So a natural way to try and correct errors is simply to repeat each transmitted word. However, it's not enough just to repeat the message once. To see this, suppose we are still trying to tell the robot arm in Example 2.2 to move DOWN,

and we transmit the instruction twice, namely 0101. If again a single error occurs (say in the first bit), and the arm receives 1101, it can certainly detect that there has been an error, since 11 is different from 01. However, there is still no way of correcting the error – that is, no way of deciding which bit is wrong – because as far as the receiver is concerned the original message *could* have been 1111 with a transmission error in the third bit. Now let's try transmitting the message *three* times: 010101. If a single error now occurs in transmission, say in the third bit so 011101 is received, not only is this error detected but it can also be corrected on a 'best of three' count for each bit, as shown:

$$\begin{array}{l} 011101 \\ \underline{\uparrow\uparrow\uparrow} \\ \text{on a majority count,} \\ \text{this bit should be 0} \end{array} \qquad (2.1)$$

Although this simple *repetition code* can correct any single transmission error, an obvious disadvantage is the need to transmit three times the actual information. This is not only expensive, but may be physically impractical – think of the problem when data are to be sent back to Earth by a space probe. Notice, as well, that since the extra bits (the repetitions) are themselves subject to error, accuracy cannot be guaranteed. However, because of the general reliability of electronic equipment, we are justified in making the assumption throughout this chapter that the probability of an error in a single bit is small, so that one transmission error is more likely than two (or more) errors. To see what this implies, let's look again at the received message 011101 in (2.1). It is certainly conceivable from the receiver's point of view that the intended message was 11 (RIGHT), that we sent 111111 and *two* transmission errors occurred in the first and fifth bits, thereby producing (2.1). However, this is less likely to have happened than our original deduction that 010101 was sent with a single error occurring in the third bit. All we aim to do, then, is to make the probability of accuracy as high as possible.

It should be stressed that at this stage we are supposing that decoding of a received word is done on the basis of direct comparison. That is, a complete list of codewords is available at the receiving end, and the received word is simply compared with these. We choose as the transmitted codeword the one which is obtained from the received word with the *smallest* number of errors – this is called *nearest-neighbour* (NN) decoding.

EXERCISE 2.2 A message to be transmitted consists of a single bit, 0 or 1 (YES or NO). The repetition code used is

Message	0	1
Codeword	000	111

For example, if 001 is received, then assuming at most a single transmission error, we deduce that 000 was transmitted, so the message is decoded as 0. By considering all the remaining seven possible received words, show that this code corrects all single errors.

A very simple but useful code is obtained by appending just one extra bit to each information message so that the overall number of ones is even – this produces the *even parity code* (the *odd parity code* similarly has an odd number of ones in each codeword). If *any single* error occurs in transmission then the number of ones in the received word will be odd, so that we detect the error and can then ask for the message to be retransmitted.

■ EXAMPLE 2.3

Consider the code in Example 2.2, and put either 0 or 1 onto the end of each word so as to make each new codeword contain an even number of ones. For example, 01 becomes 011, and the complete table of new codewords is

UP	DOWN	LEFT	RIGHT
000	011	101	110

If, say, 011 is sent and an error occurs in the second bit so that 001 is received, the error is immediately detected since 001 contains an odd number of ones.

This extra bit is called the *parity-check bit*, or simply the *check bit*, and the original bits are called the *information bits*. We shall only deal with so-called *block codes* where each codeword has the same *length*, which is the total number of symbols; this is also defined to be the length of the code itself. In Example 2.3 the codewords therefore have length 3. Codes in which the codewords have variable lengths are also used, the most famous being the *Morse code* which was particularly popular in the days before the radio transmission of speech had been invented. This code takes advantage of the relative frequencies of letters in English, so that · stands for E, and ·––– for J, for example. A snag is that it is difficult to recognize the end of a codeword, or the start of another.

■ EXAMPLE 2.4

For a code of length 7 with one even parity-check bit, suppose we wish to send the information message 110111. To make the overall parity even we append the check bit 1, and transmit the codeword 1101111. If this is the word which is received, since the overall parity is even we infer that no errors have occurred in transmission, so on dropping the check bit we correctly decode the information messages as 110111. However, if a single transmission error occurs the overall parity of the received word will be odd – for example, some possible received words containing a single error are

0101111, 1100111, 1101110

We certainly detect that an error has occurred, since each of these words has odd parity (they contain five ones). However, we cannot determine in which bit an error has occurred, so we decode the message by reporting 'Error'.

EXERCISE 2.3 For a code of length 6 with one even parity-check bit, the following words are received: 110001, 001100, 101010, 111110. Decode them, assuming at most one transmission error has occurred. For the last word, give two possible transmitted codewords which differ from this received word in only a single bit.

■ *EXAMPLE 2.5*

5 010611 18171 0

Not all codes are binary. Indeed, the example of a bar product code shown in Figure 2.1 is a *decimal* code, under the European Article Number (EAN) system. Codewords have the form

$$x_1 x_2 x_3 x_4 \ldots x_{10} x_{11} x_{12} x_{13}$$

where each of the symbols x_i can be any of the decimal digits $0, 1, 2, 3, \ldots, 8, 9$. In this code, the first two digits x_1 and x_2 are allocated to countries, with 50 belonging to the United Kingdom. The next five digits (in this example 10611) gives the manufacturer's number, and the next five (here 18171) give the unique number identifying the particular product. The last digit x_{13} is the check digit, calculated so that the *check sum*

$$x_1 + x_3 + x_5 + x_7 + x_9 + x_{11} + 3(x_2 + x_4 + x_6 + x_8 + x_{10} + x_{12}) + x_{13} \qquad (2.2)$$

is a multiple of 10. In this example we have

$$5 + 1 + 6 + 1 + 8 + 7 + 3(0 + 0 + 1 + 1 + 1 + 1) + x_{13} = 40 + x_{13}$$

so the check digit x_{13} is zero. Notice that this code detects all errors in a single digit, because if any *one* of the digits $x_1, x_2, \ldots, x_{13}$ is incorrect, the check sum (2.2) cannot be a multiple of 10.

The barcode is read by a laser scanner which works on the ratios of the widths of light and dark bars. You can see this in action at the checkout counters of most supermarkets and stores. The price of the product, however, is not part of the barcode. That information is held in the store's computer, which informs the electronic cash register of the price of each item as its barcode is scanned.

EXERCISE 2.4 Determine the check digit for the product having number 501015256020.

Another common kind of error which arises when entering numbers onto a keyboard, or reading them aloud over the telephone, is inadvertently to *transpose* (i.e. interchange) two adjacent digits. Suppose, for example, that in the product barcode in Example 2.5 x_3 and x_4 are transposed, and that $x_3 \neq x_4$ (if $x_3 = x_4$, transposition has no effect!). The change in the check sum can be calculated in the following way. In (2.2) the term x_3 is replaced by x_4, and the term $3x_4$ is replaced by $3x_3$, so the net change in the check sum is

$$-x_3 + x_4 - 3x_4 + 3x_3 = 2(x_3 - x_4) \qquad (2.3)$$

This shows that the transposition error will *not* be detected if $x_3 - x_4 = \pm 5$, since the check sum will then remain a multiple of 10. All other transposition errors will be detected, however. For example, if the barcode in Figure 2.1 was incorrectly read as 5010161181710 (the digits $x_5 = 6$ and $x_6 = 1$ being transposed) the new check sum would be

$$5+1+1+1+8+7+3(0+0+6+1+1+1)+0=50$$

which is still a multiple of 10. Hence the error of transposing the fifth and sixth digits would go undetected (here $x_5 - x_6 = 6-1=5$). If, however, the eleventh and twelfth digits were transposed, giving 5010611181170 (here $x_{11} - x_{12} = 7-1=6$), the new check sum would be

$$5+1+6+1+8+1+3(0+0+1+1+1+7)+0=52$$

which is not a multiple of 10, so the error is detected.

EXERCISE 2.5 Suppose that the factor 3 multiplying the sum of the even-numbered digits in the check sum (2.2) is replaced by 4 or 5. Explain why in both of these cases not all single-digit errors would be detected. Investigate what would happen to transposition errors.

■ *EXAMPLE 2.6*

The United States postcode, known as the 'Zip code' consists of nine decimal digits. For example, a business return envelope for SIAM (Society for Industrial and Applied Mathematics) is shown in Figure 2.4.

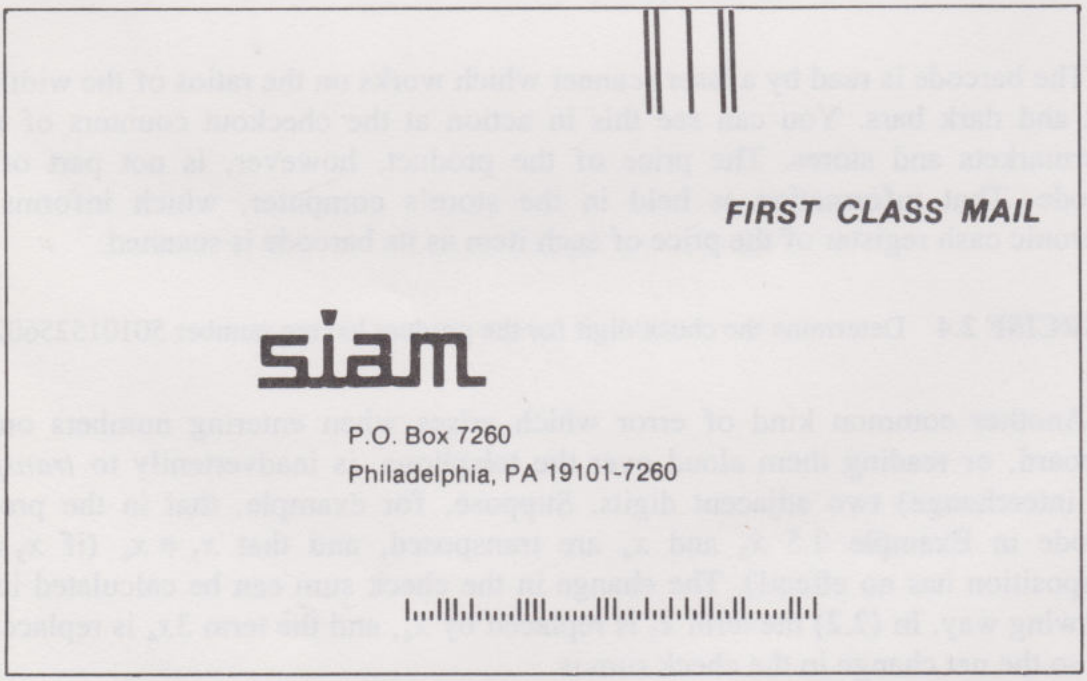


Figure 2.4

The first digit represents one of 10 geographical areas, usually a group of States, from 0 in the northeast to 9 in the far west. The next two digits identify a mail-distribution centre; the next two represent the town, or local post office. The code was first introduced in 1963, and the last four digits (7260 in Figure 2.4) were added in 1983 to facilitate computerized sorting. The first two digits of the four-digit suffix represent a delivery sector (e.g. a group of streets) and the last two narrow down the area still further, for example one floor in a large office building. The barcode shown in Figure 2.4 actually includes a tenth digit, which is the check digit. The mathematics involved in interpreting the Zip code is interesting and straightforward, and you are invited to explore this in Problem 2.1 at the end of the chapter.

It is useful at this point to introduce a special language of what are called *congruences*. If a and b are integers and their difference $a - b$ is a multiple of a third integer m (assumed positive), we say that a is *congruent to b modulo m*, and write

$$a = b(\text{mod } m)$$

In other words, there is another integer k such that

$$a - b = km \quad \text{or} \quad a = b + km$$

You are already familiar with congruences from everyday life. For example, if the hands of a conventional clock show 12 noon, then 80 minutes later the minute hand shows 20 minutes past the hour; that is, we read the minutes on a clock modulo 60. Similarly, 13 hours after noon the hour hand is against the digit 1, since we read hours modulo 12. In a similar fashion, calendars are used modulo 7 for days of the week – if 5 August is a Monday, then 12 August will also be a Monday.

■ *EXAMPLE 2.7*

We write $59 = 4(\text{mod } 11)$, since $59 - 4 = 5 \times 11$, and $37 = -3(\text{mod } 10)$ since $37 - (-3) = 4 \times 10$.

EXERCISE 2.6 What time does a conventional clock (i.e. with hands) show

(i) 17 hours after it shows 2 o'clock
(ii) 80 hours after it shows 11 o'clock
(iii) 40 hours before it shows noon?

In Example 2.5 on the product barcode, we can now say that the check sum (2.2) is required to be $0(\text{mod } 10)$.

Congruences have many properties which are the same as those for equalities. For example, if a and b are two integers such that

$$a = b(\text{mod } m) \tag{2.4}$$

then we can add any other integer c to both sides to produce

$$a + c = (b + c)(\text{mod } m)$$

since this means that $(a+c)-(b+c)=a-b$ is a multiple of m. Similarly, we can subtract c from both sides of (2.4) to give

$$a-c=(b-c)(\text{mod } m)$$

and we can multiply both sides of (2.4) by c, giving

$$ac=bc(\text{mod } m)$$

since this means $ac-bc=(a-b)c$ is a multiple of m.

More care is needed when dividing both sides of a congruence (2.4) by an integer c. For example, we have $14=8(\text{mod } 6)$, but dividing both sides by 2 gives $7=4(\text{mod } 6)$ which is not true, since $7-4$ is not a multiple of 6. However, division of both sides of (2.4) by c *is* valid provided c and m have no common factors. For example, for the congruence $14=2(\text{mod } 3)$ division throughout by 2 is now permissible since 2 and m $(=3)$ are relatively prime, so this division gives the correct congruence $7=1(\text{mod } 3)$.

Carrying out operations on integers in the way we have just described is called *modular arithmetic*. You must be thinking that we have drifted far away from our discussion of codes! In fact, what we have learnt *is* indeed relevant, as we now demonstrate.

■ EXAMPLE 2.8

Let's look again at the International Standard Book Number (ISBN), an example of which was displayed in Figure 2.2. In general an ISBN is a 10 digit codeword $x_1x_2x_3 \ldots x_9x_{10}$ which uniquely identifies a book. The first digit, x_1, denotes the country (the UK, USA and some others have 0), x_2x_3 is the publisher's number (e.g. 13 is Prentice Hall), the next six digits are the book number assigned by the publisher, and the last digit x_{10} is the check digit. In fact, there are variations: some countries are represented by more than one digit (e.g. Denmark is 87) and some publishers by more than two digits (e.g. Wiley–Interscience is 471), in which cases the assigned book number has less than six digits. The digits x_1 to x_9 can be any decimal digit from 0 to 9, but x_{10} can also take the value 10, for which the Roman numeral X is used. The check digit x_{10} is chosen so that the *check sum*, which is defined to be

$$\sum_{i=1}^{10} ix_i = x_1 + 2x_2 + 3x_3 + 4x_4 + 5x_5 + 6x_6 + 7x_7 + 8x_8 + 9x_9 + 10x_{10}$$

is $0(\text{mod } 11)$ – that is, it is a multiple of 11. Notice that we are using a congruence with $m=11$, which is a prime number (i.e. has no factors except 1 and 11). We shall study the ISBN in some detail in Section 2.5, where we'll see that the choice of (mod 11) rather than (mod 10) is crucial in endowing the code with the desirable properties of being able to detect all single-digit errors *and* all errors involving the interchange of two digits. We'll also find that modular arithmetic forms one way of constructing what are called *finite fields*.

■ *EXAMPLE 2.9*

The identification number on money orders issued by the United States Postal Service consists of 10 digits $x_1 \ldots x_{10}$ and a check digit, each digit x_i being allowed to take any value from 0 to 9. The check digit x_{11} is defined to be the remainder modulo 9 of the 10-digit number. To simplify the discussion, let's consider instead a five-digit codeword $x_1x_2x_3x_4x_5$, where the check digit x_5 is defined by the same rule, namely

$$x_1 x_2 x_3 x_4 = x_5 \pmod 9 \tag{2.5}$$

For example, if the first four digits are 5370 then $5370 = 6 \pmod 9$ so $x_5 = 6$ (since $5370 = 9 \times 596 + 6$) so the codeword is 53706. However, if the 0 is replaced by 9, the congruence becomes $5379 = 6 \pmod 9$ so that the check digit is the same, despite the error in x_4. In other words, this error in x_4 cannot be detected – the substitution of 9 for 0 does not affect the value of the check digit x_5. Clearly, the same applies if x_4 had originally been 9 and was replaced by 0. In fact, this is true for any of the first four digits – as another example, suppose x_2 was 0 and is replaced by 9, so that $x_1 0 x_3 x_4$ becomes $x_1 9 x_3 x_4$. The change in the four-digit number is 900, which is $0 \pmod 9$. By carrying on with this argument, you should be able to convince yourself that substitution of 0 by 9, or 9 by 0, in any one of x_1, x_2, x_3 or x_4 goes undetected. This does not apply to the check digit x_5, where any error *will* be detected since (2.5) will be violated.

This code is therefore not a very good one, and it's interesting to work out just how poor it is, by calculating what percentage of single errors will go undetected. First of all, we see that since each of x_1, x_2, x_3, x_4 can take one of the 10 values 0 to 9, there are a total of 10^4 possible codewords (once $x_1x_2x_3x_4$ is fixed, x_5 is determined uniquely by (2.5)). Let's count the number of words containing a single error. An error can occur in a single digit in nine possible ways. Hence there are 9×10^4 words which have the first digit incorrect, but the other four digits correct. The same applies for each of the other four digits, so in total there are $5 \times 9 \times 10^4 = 45 \times 10^4$ words containing a single error. To count the undetected single errors, consider those occurring in the first digit. There are 10^3 codewords having 0 in the first position, which if replaced by 9 go undetected; similarly, if 9 is replaced by 0, altogether 2×10^3 errors in the first digit are undetected. Since undetected errors of this type can occur in any of the first four digits, the total number is $4 \times 2 \times 10^3 = 8 \times 10^3$. Hence the detection rate for single errors is

$$\frac{\text{no. of detected errors}}{\text{total no. of errors}} = \frac{45 \times 10^4 - 8 \times 10^3}{45 \times 10^4} = \frac{442}{450}$$

or approximately 98.2%. Thus about 1.8% of single errors are undetected.

This code is much worse when it comes to errors where two consecutive digits are transposed. Suppose $x_1x_2x_3x_4x_5$ is incorrectly recorded as $x_2x_1x_3x_4x_5$. The difference between the two four-digit numbers is

$$\begin{aligned} x_1x_2x_3x_4 - x_2x_1x_3x_4 &= x_1 10^3 + x_2 10^2 + x_3 10 + x_4 - (x_2 10^3 + x_1 10^2 + x_3 10 + x_4) \\ &= 900(x_1 - x_2) \end{aligned}$$

which is obviously divisible by 9, irrespective of the values x_1 and x_2. Hence the

check digit determined from (2.5) is unaltered, so this error goes undetected. You should again convince yourself that this happens for transposition errors involving x_2 and x_3, or x_3 and x_4. However, if x_4 and x_5 are transposed then the error will be detected because the congruence (2.5) will not be satisfied (see Exercise 2.7 below). Hence the only transposition errors which are detected are those involving the check digit.

EXERCISE 2.7 Show that the congruence (2.5) is the same as

$$x_1 + x_2 + x_3 + x_4 = x_5 \pmod 9$$

(Hint: write $x_1x_2x_3x_4 = 10^3x_1 + 10^2x_2 + 10x_3 + x_4$.)

Hence show that if $x_1x_2x_3x_4x_5$ (with $x_4 \neq x_5$) is erroneously replaced by $x_1x_2x_3x_5x_4$ then x_4 is not the check digit for $x_1x_2x_3x_5$, so this transposition error is detected. (For a generalization, see Problem 2.2 at the end of the chapter.)

EXERCISE 2.8 The identity number on machine-readable passports is a seven-digit codeword $x_1x_2x_3x_4x_5x_6x_7$. The first six digits are the date of birth in the form

x_1x_2	x_3x_4	x_5x_6
day	month	year

and the check digit x_7 is chosen so as to satisfy

$$x_7 + 7(x_1 + x_4) + 3(x_2 + x_5) + x_3 + x_6 = 0 \pmod{10}$$

Confirm that this code detects all single-digit errors. Investigate what happens to transposition errors, giving careful attention to the transposition of x_4 and x_5, and of x_5 and x_6.

2.2 HAMMING DISTANCE

Let's go back to the situation in Example 2.2, where we had four messages 00, 01, 10, 11 to be sent. The difficulty was that if a single transmission error occurs then an incorrect message is received. This is because the messages are 'too close together' – an error in just 1 bit changes one word into another. We can make this idea precise by defining the Hamming distance, first suggested by R.W. Hamming in 1950. If

$$a = a_1a_2 \ldots a_n, \quad b = b_1b_2 \ldots b_n$$

are two words each of length n, then the *Hamming distance* $\delta(a, b)$ between them is the number of places in which they differ.

■ *EXAMPLE 2.10*

For a binary code of length 5 we have

$$\delta(01010, 11001) = 3$$

since $a = 01010$ and $b = 11001$ differ in the first, fourth and fifth places. Similarly, for a decimal code of length 4 we have $\delta(9172, 8272) = 2$, since the words differ in the first and second places.

If we think of the two words a and b as being 'points' in space, then in order to 'travel' from one word to the other we have to change precisely $\delta(a, b)$ digits – this number of changes is what we call the Hamming distance. More precisely, this distance justifies its name since it has exactly the same three mathematical properties as the normal concept of geometrical distance between two points in space. Two of these properties are obvious: first, the distance between a and b is zero if and only if a and b coincide, that is

$$\delta(a, b) = 0, \quad \text{if and only if } a = b$$

and otherwise $\delta(a, b) > 0$.

Secondly, the distance from one word to another doesn't depend on the 'direction' of travel, that is

$$\delta(a, b) = \delta(b, a), \quad \text{for all } a \text{ and } b$$

The third property is called the *triangle inequality* for the following reason. Suppose we have three points A, B, C forming a plane triangle, as shown in Figure 2.5, with $d(\mathrm{A}, \mathrm{B})$ denoting the geometrical distance between A and B, and similarly for the other two sides.

It is clear that

$$d(\mathrm{A}, \mathrm{B}) \leqslant d(\mathrm{A}, \mathrm{C}) + d(\mathrm{C}, \mathrm{B})$$

The corresponding result for the Hamming distance looks just the same: for any third codeword $c = c_1 c_2 \ldots c_n$ we have

$$\delta(a, b) \leqslant \delta(a, c) + \delta(c, b) \tag{2.6}$$

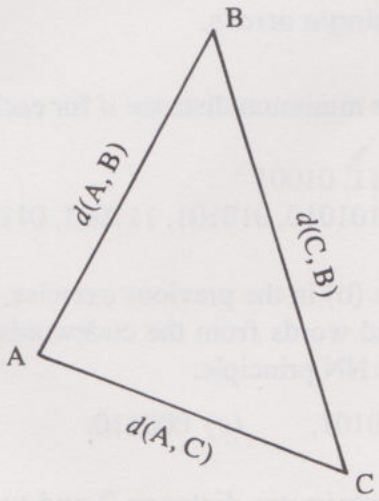


Figure 2.5

To verify (2.6), we must realize that another way of thinking of the Hamming distance $\delta(a, b)$ is that it is the *smallest* number of changes to digits of the codeword a needed to produce b. For example, we saw in Example 2.10 that $\delta(9172, 8272) = 2$, and the first two digits of $a = 9172$ need to be changed in order to obtain $b = 8272$. In general, in order to change a into b we can first change it to c, which requires $\delta(a, c)$ changes, and then go from c to b, requiring a further $\delta(c, b)$ changes. Since $\delta(a, b)$ is the smallest number of changes needed to go from a to b, it can't be bigger than the sum of $\delta(a, c)$ and $\delta(c, b)$ – so we have proved (2.6).

The use of the term 'nearest-neighbour' (NN) decoding, which we introduced earlier, can now be reinterpreted: we choose as the most likely transmitted word the one which is *nearest* (as measured by the Hamming distance) to the received word.

You should by now have some feeling that a crucial parameter affecting the properties of a code is going to be the overall closeness, in the Hamming distance sense, of codewords. This is measured by the *minimum distance d*, which is the smallest value of all the distances between all possible pairs of (different) codewords.

■ EXAMPLE 2.11

For the code in Example 2.2 we can very easily work out that

$$\delta(00, 01) = 1, \quad \delta(00, 10) = 1, \quad \delta(00, 11) = 2$$

$$\delta(01, 10) = 2, \quad \delta(01, 11) = 1, \quad \delta(10, 11) = 1$$

This shows that the minimum distance is 1, and explains why the code is useless. For *any* code which has minimum distance 1, there will (by definition) be at least one pair of codewords a and b for which $\delta(a, b) = 1$. If a is sent, and an error occurs in that one particular digit in which a differs from b, then the codeword b will be received, which will be assumed correct – there is no way of telling that an error has occurred. Hence a code whose minimum distance is 1 cannot even detect all single errors.

EXERCISE 2.9 Determine the minimum distance d for each of the following binary codes:

(a) $C = \{1000, 1011, 0100\}$
(b) $C = \{000000, 101010, 010101, 111001, 011110\}$.

EXERCISE 2.10 For the code (b) in the previous exercise, determine the distances of each of the following received words from the codewords, and hence decode each of the received words using the NN principle:

(a) 100010, (b) 000101, (c) 000110.

However, if a code has minimum distance 2 and we transmit a codeword a, any single error will result in a received word which is *not* a codeword (since all

codewords differ from a in at least two places). Hence any single error will always be detected. What about error correction? Unfortunately, if $d=2$ there will always be at least one uncorrectable single error. To see this, suppose for simplicity that the code has length 4, and consider two codewords

$$a=a_1a_2a_3a_4, \quad b=a_1b_2a_3b_4$$

with $b_2 \neq a_2$, $b_4 \neq a_4$, so that $\delta(a, b)=2$. Suppose a is transmitted and a single error occurs in a_2, causing it to become b_2, so the received word is $c=a_1b_2a_3a_4$. Clearly $\delta(c, a)=1$ and $\delta(c, b)=1$. Thus c cannot be a codeword, since all codewords are distance 2 at least apart, and hence an error is detected. However, there is no way of deciding by the NN principle whether a or b was sent, so the error cannot be corrected. You should have no difficulty in seeing how the argument still applies whatever the wordlength.

We are beginning to see how we can quantify our intuitive feeling that the further apart codewords are (in the sense of Hamming distance), the better will be the code from the point of view of coping with errors. Let's explore this with an example having $d=3$, before trying to get a general result.

■ *EXAMPLE 2.12*

We return to the repetition code, introduced in Example 2.2. Each 2 bit message was transmitted three times, giving the codewords

$$a_1=000000, \quad a_2=010101, \quad a_3=101010, \quad a_4=111111$$

To find the minimum distance it is convenient to display the distances between pairs of codewords in the following tabular form:

	a_1	a_2	a_3	a_4
a_1	–	3	3	6
a_2	3	–	6	3
a_3	3	6	–	3
a_4	6	3	3	–

The smallest number appearing in the table is 3, which is therefore the minimum distance for this repetition code. Notice that the table is symmetrical relative to the principal diagonal (top left corner to bottom right) – that is, the numbers in the first row are the same as those in the first column, the second row is identical to the second column, and so on. This is because of the property $\delta(a, b)=\delta(b, a)$, so a table of distances for any code will always be symmetric. Only the upper triangular part need therefore be recorded.

We saw that this repetition code will always correct a single transmission error, and in fact this is true for any code having minimum distance 3. This is a special case of the second part of the following important result.

Theorem 2.1

Let C be a code having minimum distance d.

(i) C will *detect* e errors using the NN principle provided

$$d \geqslant e+1 \tag{2.7}$$

(ii) C will *correct* e errors using the NN principle provided

$$d \geqslant 2e+1 \tag{2.8}$$

Proof

(i) From the definition of minimum distance, all codewords differ in at least d places, i.e. by (2.7), in at least $e+1$ places. Therefore, if a codeword is transmitted and at most e transmission errors occur, then the received word cannot be a codeword. Hence these errors are detected.

(ii) Suppose a codeword a is sent and e errors occur, so that a word c is received for which

$$\delta(a, c) = e \tag{2.9}$$

Let b be *any* other codeword different from a, so we have $d \leqslant \delta(a, b)$, and therefore by (2.8)

$$2e+1 \leqslant \delta(a, b) \tag{2.10}$$

Substitute (2.9) and (2.10) into the triangle inequality (2.6) to obtain

$$\begin{aligned} 2e+1 &\leqslant \delta(a, b) \\ &\leqslant \delta(a, c) + \delta(c, b) \\ &\leqslant e + \delta(c, b) \end{aligned}$$

for which we get $\delta(c, b) \geqslant e+1$. Thus b is a distance greater than e from the received word c – so a is the *only* codeword within a distance e from c. The NN principle therefore correctly decodes the received word to produce the original message a.

Theorem 2.1 can be interpreted as saying that a code with minimum distance d can be used either to detect $d-1$ errors, or to correct $\frac{1}{2}(d-1)$ errors (if d is odd) and $\frac{1}{2}(d-2)$ errors (if d is even). This agrees with what we have already discovered for $d = 1, 2, 3$. Notice that the proof is not in any way restricted to binary codes. It's interesting that the Mariner 9 code mentioned in Section 2.1 had length 32, 26 check bits and minimum distance 16, and so could correct up to seven errors.

■ ***EXAMPLE 2.13***

Consider the code

$C = \{00110, 10001, 01011, 11100\}$

It is left as an exercise for you to check that the minimum distance is $d = 3$. The theorem tells us that this code can either correct one error, or detect two.

For example, suppose that 00011 is received. The distances from the four codewords are respectively 2, 2, 1, 5. Thus the received word is nearest to the third codeword, so we decide by the NN principle that 01011 was transmitted, with a single error in transmission (in the second bit).

However, if 01101 is received then the respective distances from the four codewords are 3, 3, 2, 2 so we cannot decide by the NN principle whether the third or fourth codeword was transmitted. However, we have detected that there are two errors – for example, 01101 could have come from 11100 with errors in the first and last bits.

EXERCISE 2.11 What is the smallest possible minimum distance that a code must have in order to correct two errors? How many errors will it detect?

EXERCISE 2.12 A code has minimum distance 3. Show that it is not possible to correct all single errors *and* detect all double errors. That is, show that there exist codewords a and b and a received word c, such that c comes from a via one error, and from b via two errors.

2.3 LINEAR BINARY CODES

We now consider a binary code C consisting of codewords $a = a_1 a_2 a_3 \ldots a_n$, where each element a_i is 0 or 1. Define the *sum* of two codewords as $c = a + b$, where $c_i = a_i + b_i$, $i = 1, 2, \ldots, n$, that is we add the bits term by term, and apply the following rules:

$$0+0=0, \quad 1+0=1, \quad 0+1=1, \quad 1+1=0 \tag{2.11}$$

These rules are in fact addition modulo 2 (which can be interpreted as even + even = even, odd + even = odd, odd + odd = even, where '0' stands for an even number and '1' for an odd number).

■ ***EXAMPLE 2.14***

If 1101 and 1001 are two codewords in a code of length 4, then their sum is

$$1101 + 1001 = 0100$$

Note that this is *not* binary arithmetic. If 1101 and 1001 were regarded as binary numbers instead of codewords, then their sum would be

$$\begin{array}{r} 1101 \\ 1001 \\ \hline 10110 \end{array}$$

where we 'carry over' as in familiar decimal arithmetic.

A *linear* code C is one in which the sum of *any* two codewords is also a codeword: that is, if a, b are in C then so is $a+b$. In particular, taking $b=a$ shows that any linear code always contains the *zero word* $\mathbf{0}$ (consisting only of zero bits) since the ith bit of $a+a$ is either

$$a_i + a_i = 1 + 1 = 0 \quad \text{or} \quad a_i + a_i = 0 + 0 = 0$$

■ EXAMPLE 2.15

(a) The code $C=\{00, 10, 01, 11\}$ is linear because all possible sums of codewords are also codewords, that is

$$10+01=11, \quad 10+11=01, \quad 01+11=10$$
$$00+10=10, \quad 00+01=01, \quad 00+11=11$$

(b) The code $C=\{0000, 0101, 1011\}$ is *not* linear, because

$$0101+1011=1110$$

which is not a codeword.

EXERCISE 2.13 Determine whether each of the following sets of codewords forms a linear code:

(a) 000, 110, 100
(b) 000, 100, 011, 111
(c) 00000, 01110, 10111, 11001.

An important reason for using linear codes is that there is an easy way of calculating the minimum distance. To find out what that is, we first need to define the *weight* $w(a)$ of the codeword a as the number of ones in a. For example, $w(1101)=3$. If a and b are any two codewords belonging to a linear code, then in their sum $c=a+b$ the ith bit $c_i=a_i+b_i$ is 1 if a_i and b_i differ, and $c_i=0$ if a_i and b_i are the same. Hence the weight of c is simply equal to the number of places in which a and b differ – in other words, we have proved that

$$w(a+b)=\delta(a, b) \tag{2.12}$$

In particular, if $b = \mathbf{0}$ in (2.12) when we get

$$w(a) = \delta(a, \mathbf{0}) \tag{2.13}$$

The key result can now be established.

Theorem 2.2

For any linear code C, its minimum distance d is equal to the smallest non-zero weight of codewords, that is

$$\begin{aligned} d &= \min_{a \neq 0} w(a) \\ &= w_{\min}, \text{ say} \end{aligned} \tag{2.14}$$

Proof

The argument is quite ingenious, and consists of showing that the minimum distance d is not greater than $w_{\min}$, and also is not less than $w_{\min}$; hence d is 'sandwiched' on both sides by $w_{\min}$, and so must be equal to it.

First, let c be a codeword for which $w(c) = w_{\min}$ (by definition, there must be at least one such c). Since $\mathbf{0}$ is also a codeword, by definition of minimum distance we have $d \leqslant \delta(c, \mathbf{0})$. From (2.13) we have $\delta(c, \mathbf{0}) = w(c)$, so combining these two facts produces $d \leqslant w_{\min}$ (i.e. d is not greater than $w_{\min}$).

Next, let f and g be two codewords which are distance d apart, i.e. $d = \delta(f, g)$. However, the sum $f + g$ is also a codeword so from (2.12) we have

$$\delta(f, g) = w(f + g) \geqslant w_{\min}$$

Hence $d \geqslant w_{\min}$, and this is the second half of the 'sandwich' argument which shows that $d = w_{\min}$.

Armed with Theorem 2.2, we can now see why it is easy to determine the minimum distance for a linear code. Instead of having to compute $\delta(a, b)$ for all possible pairs of codewords a and b, and then finding the minimum of these distances, we simply have to compute the minimum of the weights of all the (non-zero) codewords.

■ ***EXAMPLE 2.16***

It is left as an easy task for the reader to check that the code

$$C = \{00000, 01110, 10001, 11111\}$$

is linear. By inspection, the non-zero weights are 3, 2, 5 respectively, so the minimum distance is 2.

EXERCISE 2.14 Verify that each of the following sets constitutes a linear code, and find the minimum distance:

(a) 00000000, 10101010, 01010101, 11111111
(b) 00000, 00111, 01011, 01100, 10011, 10100, 11000, 11111.

2.4 MATRIX REPRESENTATION

We would now like to be able to construct linear codes which correct all single errors, and to do this we need to utilize matrix notation. Recall from (1.59) how to multiply a matrix by a vector. For example, if

$$A = \begin{bmatrix} a_1 & a_2 & a_3 \\ a_4 & a_5 & a_6 \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

then their product is

$$Ab = \begin{bmatrix} a_1b_1 + a_2b_2 + a_3b_3 \\ a_4b_1 + a_5b_2 + a_6b_3 \end{bmatrix}$$

We shall only need *binary matrices*, which have elements which are either 0 or 1, and the arithmetic is carried out modulo 2, according to the rules in (2.11).

■ *EXAMPLE 2.17*

Using the multiplication rule above we have

$$\begin{bmatrix} 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1+0+0+1+1 \\ 0+0+1+1+0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

EXERCISE 2.15 Compute the following products, using modulo 2 arithmetic:

(a)

$$\begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

(b)

$$\begin{bmatrix} 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

Let $a = a_1 a_2 \ldots a_n$ and $b = b_1 b_2 \ldots b_n$ be *any* two codewords belonging to a code we are trying to construct, and suppose H is a binary matrix with n columns. Let's write the codewords as column vectors, for which we shall use the notation

$$a' = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_n \end{bmatrix}, \quad b' = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{bmatrix}$$

This is not to be confused with the transpose notation $[a_1, a_2, \ldots, a_n]^T$ which turns a *row vector* into a column vector.

The crucial ploy is to choose the matrix H so that codewords satisfy the equations

$$Ha' = 0, \quad Hb' = 0$$

Clearly if $c = a + b$, then

$$Hc' = H(a' + b') = Ha' + Hb' = 0$$

which shows that c is also a codeword. But since c is the sum of a and b, this means that the way we have set things up ensures the code is linear. That is to say, we have characterized our linear binary code as the set of all codewords $x = x_1 x_2 x_3 \ldots x_n$ which satisfy $Hx' = 0$. The matrix H is called the *parity-check matrix*, or simply the *check matrix*.

■ *EXAMPLE 2.18*

Suppose that we take as check matrix

$$H = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

To determine the codewords $x = x_1 x_2 x_3$ for which H is the check matrix, the simple-minded approach is to take all the possible binary words of length 3 and see which of them satisfy $Hx' = 0$. Since each x_i can be 0 or 1, there are $2^3 = 8$ possible words, namely

000, 100, 010, 001, 101, 110, 011, 111

By direct multiplication we get

$$H\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = 0, \quad H\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad H\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad H\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$H\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad H\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad H\begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad H\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Only the first and last of these products give the zero vector, so we conclude that the only two codewords defined by this particular H are 000 and 111. Not a very interesting code – we can only send the messages 'yes' and 'no'!

A more systematic way of finding the codewords is illustrated by the next example.

■ *EXAMPLE 2.19*

Let's find all the codewords for the linear code determined by the check matrix

$$H = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix} \tag{2.15}$$

This requires us to solve

$$H \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

which when written out in full becomes

$$\begin{aligned} x_1 + x_4 &= 0 \\ x_2 + x_3 + x_4 &= 0 \end{aligned} \tag{2.16}$$

These are called the *check equations*. The first one gives

$$x_1 = -x_4 = x_4 \tag{2.17}$$

Notice that we can write $-x_4 = x_4$ because with arithmetic modulo 2, as expressed in (2.11), then

$$1 + 1 = 0, \quad 0 + 0 = 0$$

which is equivalent to stating that $-1 = 1$, $-0 = 0$. Thus for *any* binary x we have $-x = x$, and so the second equation in (2.16) gives

$$x_2 = -x_3 - x_4 = x_3 + x_4 \tag{2.18}$$

We have now expressed x_1 and x_2 in terms of x_3 and x_4, which can be regarded as the two independent variables. Since x_3 and x_4 can take the values 0 or 1, there are four possibilities which can be represented in tabular form as follows:

x_1	x_2	x_3	x_4
0	0	0	0
1	1	0	1
0	1	1	0
1	0	1	1

For each pair of values of x_3 and x_4, the corresponding values of x_1 and x_2 are obtained from (2.17) and (2.18). For example, when $x_3 = 1$, $x_4 = 1$ then $x_1 = 1$, $x_2 = 1 + 1 = 0$. The four codewords are therefore 0000, 1101, 0110, 1011.

EXERCISE 2.16 Determine all the codewords for the linear code having check matrix

$$H = \begin{bmatrix} 1&0&0&0&1&1&1 \\ 0&1&0&0&0&1&1 \\ 0&0&1&0&1&1&1 \\ 0&0&0&1&0&0&1 \end{bmatrix}$$

We are now ready to develop a general procedure for constructing a linear code. The equations (2.16) in Example 2.19 were easy to solve because x_1 appeared only in the first one, and x_2 only in the second one. This is because the first two columns of H in (2.15) consist of I_2, the 2×2 unit matrix. Similarly, for the matrix H in Exercise 2.16 the first four columns consist of I_4. In general, we can take as check matrix the $r \times n$ matrix

$$H = [I_r \quad A] \tag{2.19}$$

which has r rows and n columns. In (2.19) I_r is the $r \times r$ *unit matrix* (defined in Section 1.4, Chapter 1)

$$I_r = \begin{bmatrix} 1 & & & \\ & 1 & & \mathbf{0} \\ & & \ddots & \\ & \mathbf{0} & & 1 \end{bmatrix}$$

having ones on the principal diagonal and zeros elsewhere, and A is an arbitrary $r \times (n-r)$ binary matrix whose element in row i, column j, we denote by a_{ij}. Codewords $x_1 x_2 \ldots x_n$ have length n and satisfy the condition $Hx' = 0$. For simplicity let's write out the case $r = 3$, $n = 5$:

$$\begin{bmatrix} 1&0&0&a_{11}&a_{12} \\ 0&1&0&a_{21}&a_{22} \\ 0&0&1&a_{31}&a_{32} \\ \multicolumn{3}{c}{I_3} & \multicolumn{2}{c}{A} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = 0$$

We can write the check equations as

$$\begin{aligned} x_1 \quad &+ a_{11}x_4 + a_{12}x_5 = 0 \\ x_2 \quad &+ a_{21}x_4 + a_{22}x_5 = 0 \\ x_3 \quad &+ a_{31}x_4 + a_{32}x_5 = 0 \end{aligned}$$

As before, because we are using arithmetic modulo 2 we can rewrite these as

$$\begin{aligned} x_1 &= a_{11}x_4 + a_{12}x_5 \\ x_2 &= a_{21}x_4 + a_{22}x_5 \\ x_3 &= a_{31}x_4 + a_{32}x_5 \end{aligned}$$

or in matrix notation

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = A \begin{bmatrix} x_4 \\ x_5 \end{bmatrix} \tag{2.20}$$

The variables x_4 and x_5 are the independent or *information bits* which we can choose arbitrarily (depending on the message to be transmitted). The *check bits* x_1, x_2, x_3 are then determined uniquely from (2.20). The general version of (2.20) is

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_r \end{bmatrix} = A \begin{bmatrix} x_{r+1} \\ x_{r+2} \\ \vdots \\ x_n \end{bmatrix} \tag{2.21}$$

and there are $n-r$ information bits $x_{r+1}, x_{r+2}, \ldots, x_n$ and r check bits $x_1, x_2, \ldots, x_r$. Since each bit is 0 or 1, there are 2^{n-r} different ways of choosing the values of the information bits, so there are 2^{n-r} codewords in total. For this reason the quantity $k = n - r$ is called the *dimension* of the code. Notice that if the order of the columns of H in (2.19) is altered, then this simply alters the order of the bits in the codewords in the corresponding way. In particular, some books use $H = [A \quad I_r]$.

■ EXAMPLE 2.20

(a) Go back to Example 2.19. We see that in (2.15)

$$= [I_2 \; A], \quad A = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$$

and $n = 4$, $r = 2$, so the code has dimension 2 and $2^2 = 4$ codewords. From (2.21) we obtain

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_3 \\ x_4 \end{bmatrix}$$

giving the expressions (2.17) and (2.18).

(b) Let's determine the codewords for the linear code having check matrix

$$H = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 \end{bmatrix}$$

$$I_3 \qquad A$$

Here $r = 3$, $n = 7$ and the check equations (2.21) are

$$\begin{aligned} x_1 &= x_4 + x_6 + x_7 \\ x_2 &= x_4 + x_5 + x_6 \\ x_3 &= x_4 + x_5 + x_7 \end{aligned} \tag{2.22}$$

The dimension of the code is $7-3=4$, so there are $2^4=16$ codewords. As in Example 2.19, for each possible set of values of the information bits the values of the check bits given by (2.22) can be conveniently expressed in tabular form, and six of the codewords are listed below:

Check bits			Information bits			
x_1	x_2	x_3	x_4	x_5	x_6	x_7
0	0	0	0	0	0	0
1	1	1	1	0	0	0
0	1	1	0	1	0	0
1	1	0	0	0	1	0
1	0	1	0	0	0	1
1	0	0	1	1	0	0

EXERCISE 2.17 Determine the remaining 10 codewords in the preceding example by using (2.22). Verify that the codewords satisfy $Hx'=0$.

Our next task is to investigate the error detection and correction properties of a code in terms of its check matrix. Recall that for a code to be able to *detect* all single errors it must have a minimum distance d of at least 2. In view of Theorem 2.2, this means that there must be no codewords of weight 1, since d is equal to the minimum of all the weights of (non-zero) codewords. Suppose e is a word of weight 1, and so has just a single non-zero bit in (say) the ith position. Hence for e *not* to be a codeword, we must have $He' \neq 0$. However, because e' is a column vector with a single 1 in the ith element, the vector He' is just the ith column of H, and this must therefore not equal zero for any value of i from 1 to n. We have therefore proved the following theorem.

Theorem 2.3

H is the check matrix for a single-error-detecting linear binary code if and only if it does *not* contain a zero column.

■ *EXAMPLE 2.21*

If the check matrix is

$$H=\begin{bmatrix}1&0&1&0\\0&1&0&0\end{bmatrix}$$

then because the *fourth* column of H is zero, a codeword of weight 1 is 0001 (the 1 is in the fourth place). Hence $d=1$, and the code does not detect single errors.

EXERCISE 2.18 Theorem 2.3 states that *any* $r \times n$ binary matrix without a column of zeros will be the check matrix for a single-error-detecting code. Suppose that H is taken to be in the form (2.19), and write down a suitable H with four check bits and three information bits. Determine all the codewords and the minimum distance (notice that this can turn out to be *greater* than 2).

Now let's move on to *correction* of single errors. We recall that in this case we must have $d \geqslant 3$, so that there must be no codewords having weight 2; that is, for any word f having exactly two non-zero bits in positions i and j (say) we must have $Hf' \neq 0$. But the product Hf' in this case is equal to $h_i + h_j$, the sum of the ith and jth columns of H. For example, if $n=5$, $i=2$, $j=4$ we would get

$$Hf' = H\begin{bmatrix}0\\1\\0\\1\\0\end{bmatrix} = h_2 + h_4$$

We therefore require $h_i + h_j \neq 0$, or $h_i \neq h_j$ (since $-h_j = h_j$). Clearly the condition of Theorem 2.3 for $d \geqslant 2$ must also hold, so we have proved the following theorem.

Theorem 2.4

H is the check matrix for a single-error-correcting (s.e.c.) linear binary code if and only if no two columns of H are equal, and no column is zero.

■ *EXAMPLE 2.22*

It is now very simple to write down a check matrix which will produce an s.e.c. code. For example, if there are three check bits and three information bits then $n = 3+3 = 6$, and a suitable check matrix in the form (2.19) is

$$H = \begin{bmatrix}1&0&0&1&1&0\\0&1&0&1&1&1\\0&0&1&1&0&1\end{bmatrix}$$

This choice is not unique; for example, any of the columns could be replaced by

$$\begin{matrix}1\\0\\1\end{matrix}$$

The *only* requirement is that all the columns are non-zero and different from each other.

If, however, we consider

$$H=\begin{bmatrix}1&0&0&1&1&0\\0&1&0&1&0&1\\0&0&1&1&0&1\end{bmatrix}$$

then we see that the *first* and *fifth* columns are identical. Hence 100010 is a codeword of weight 2, confirming that this second matrix H cannot produce an s.e.c. code since $d<3$.

EXERCISE 2.19 Confirm that

$$H=\begin{bmatrix}1&0&0&1&0\\0&1&0&1&1\\0&0&1&1&0\end{bmatrix}$$

cannot be used as the check matrix for an s.e.c. code by (a) writing down a codeword of weight 2, and (b) determining two non-zero codewords which are distance 2 apart.

Let's now look more closely into the structure of check matrices which produce s.e.c. codes. If there are only two check bits then the only possible check matrix satisfying the conditions in Theorem 2.4 is

$$H=\begin{bmatrix}1&0&1\\0&1&1\end{bmatrix} \tag{2.23}$$

(apart from a permutation of the columns, which as we have remarked earlier merely permutes the order of the bits). If $r=3$, then we have from (2.19)

$$H=[I_3 \quad A]$$

To satisfy the conditions of Theorem 2.4 we must exclude the zero column and the columns of I_3 from A, so the only possibilities for the columns of A are

$$\begin{matrix}1&1&0&1\\1&0&1&1\\0&1&1&1\end{matrix}$$

By selecting one, two, three or all four of these columns we produce codes of lengths 4, 5, 6 or 7 respectively. In general, with r check bits and length n we have from (2.19)

$$H=\underset{r}{[I_r}\quad \underset{(n-r)}{A}\quad]r \tag{2.24}$$

There are in total 2^r possible columns to select for A, since each of the r elements in a column can be 0 or 1. However, in order to obtain the s.e.c. property we must exclude the zero column and the r columns of I_r from A. We are therefore left with at most 2^r-r-1 columns for A. Looking at (2.24), we see that

$$n-r\leqslant 2^r-r-1$$

showing that the length of codewords satisfies the condition $n\leqslant 2^r-1$. When equality holds the code is called *perfect*.

■ ***EXAMPLE 2.23***

When $r=2$ we have $n \leqslant 2^2-1=3$, as is apparent in (2.23). When $r=3$ we get $n \leqslant 2^3-1=7$, which again confirms the discussion above.

EXERCISE 2.20 For linear s.e.c. codes, how many check bits are needed with (a) 20 and (b) 32 information bits?

EXERCISE 2.21 Using a suitable check matrix, list the codewords for an s.e.c. code with two information bits and the smallest possible number of check bits.

EXERCISE 2.22 A first-year class in the School of Humorous Studies contains 59 students, and it is decided to assign to each an identity number in the form of a binary word.

(a) What is the least possible number of information bits of a linear code used for this purpose? (It is assumed that some codewords will be left over unused.)
(b) If the code must be capable of correcting all single errors, find the least possible length of codewords.
(c) Write down any suitable check matrix for such a code.

We now know how to construct an s.e.c. code for given numbers r and k of check bits and information bits respectively by choosing any suitable check matrix H in the form (2.24). In order to encode a given information message, we obtain the r check bits from (2.21). However, to decode we have so far simply compared a received word m, say, with the set of all codewords. By appealing to the NN principle we then select the codeword nearest to m as that most likely to have been transmitted. This is a tedious procedure, and in fact a much simpler decoding procedure can be developed using the check matrix H. Suppose that the received message is $m=c+e$, where c is a codeword, and e represents a single error in the ith bit, so

$$e=00\ldots\underset{\substack{\uparrow\\ i}}{1}0\ldots 0$$

By definition of the check matrix we have $Hc'=0$, so

$$\begin{aligned} Hm' &= H(c'+e') \\ &= He' \end{aligned}$$

However, because the column vector e' formed from e has a single 1 in the ith element, we saw earlier that He' is just the ith column of H. We have therefore established for our s.e.c. code the following theorem.

Theorem 2.5

If a single error occurs in transmission then Hm' is equal to some column (the ith, say) of H; and the error is in the ith bit.

Because of its role in determining the error, the vector Hm' formed by multiplying the check matrix by the vector of the received word is called the *syndrome* of m (derived from the medical usage of this word where it means 'symptom').

Syndrome decoding algorithm

Step 1 Compute $s = Hm'$.
Step 2 If $s = 0$, assume m is a codeword and no transmission error has occurred.
Step 3 If $s = i$th column of H, a single transmission error occurred in the ith bit.
Step 4 If $s \neq 0 \neq i$th column of H then more than one error occurred in transmission.

■ *EXAMPLE 2.24*

Consider the matrix

$$H = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

which is the check matrix for an s.e.c. code, since it satisfies the conditions of Theorem 2.4: all columns are non-zero and different from each other. Suppose a received word is $m = 11110$; then the syndrome is

$$s = H\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

which is the second column of H. We deduce from Step 3 of the algorithm that there is an error in the second bit, so the correctly decoded message is $c = 10110$. It is easy to check that $Hc' = 0$, verifying that c is indeed a codeword.

Suppose that a second received word is $m = 00111$. In this case the syndrome is

$$s = H\begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$$

which is *not* a column of H, so by Step 4 we conclude that more than one error has occurred in transmission. In fact, as the reader can check, in this case possible transmitted codewords are 10110, with errors in the first and last places; or 01011 with errors in the second and third places; or 11101 with errors

in the first, second and fourth places. Our decoding algorithm cannot determine the transmitted codeword in this case.

EXERCISE 2.23 Consider the code having check matrix

$$H = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

(a) A codeword is transmitted, and a single error occurs in the second bit. What is the syndrome?
(b) Decode each of the received messages: 11110, 11101, 10111.

EXERCISE 2.24 Consider the code with check matrix

$$H = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

(a) Decode the received words 111001, 110111.
(b) Show that if 111111 is received then more than one error has occurred in transmission. Find two possible codewords which could have been transmitted with two errors occurring, and a codeword which could have been transmitted with three errors.

EXERCISE 2.25 If a codeword is transmitted, and errors occur in bits i and j, show that the syndrome is the sum of columns i and j of the check matrix.

EXERCISE 2.26 If H is not the check matrix for an s.e.c. code, then of course the decoding algorithm breaks down. Verify this by considering the check matrix

$$H = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

and a received word 01110.

2.5 HAMMING CODES

An important family of s.e.c. codes was discovered by the American mathematician and computer scientist Hamming around 1950.

■ *EXAMPLE 2.25*

Suppose we wish to construct a check matrix H for an s.e.c. code with length $n = 6$ and three check bits. One natural way of selecting columns of H so that

they satisfy the conditions of Theorem 2.4 is to write down the binary number representation of the integers 1 to 6, namely

$$\begin{array}{cccccc} 001, & 010, & 011, & 100, & 101, & 110 \\ 1 & 2 & 3 & 4 & 5 & 6 \end{array}$$

Obviously these binary numbers are all different from each other and none is zero. Therefore, if we write them as the columns of a matrix

$$H=\begin{bmatrix} 0&0&0&1&1&1 \\ 0&1&1&0&0&1 \\ 1&0&1&0&1&0 \end{bmatrix} \tag{2.25}$$

this does indeed comply with the requirements in Theorem 2.4, and so is the check matrix for an s.e.c. code. Notice that H in (2.25) no longer has the standard form (2.24) where the first three columns are those of the 3×3 unit matrix I_3. However, this doesn't matter – it simply means that the check bits are in the positions where the columns of I_3 appear in (2.25), namely positions 1, 2 and 4.

Recall (Theorem 2.5) that if m is a received word and the syndrome $s=Hm'$ is equal to the ith column of H, this shows that there is a single error in the ith bit. However, because of the way we have constructed H, its ith column is just the number i in binary form, so the syndrome s itself has given us the bit which is in error, *without having to compare s with the columns of H.* This is the clever idea behind Hamming codes, which makes them particularly easy to decode. For example, if $m=101001$ then using the matrix H in (2.25), the syndrome is

$$s=H\begin{bmatrix} 1\\0\\1\\0\\0\\1 \end{bmatrix}=\begin{bmatrix} 1\\0\\0 \end{bmatrix}$$

This corresponds to 100 (we use the notation $s\sim 100$) which is the binary representation of 4. Hence the error is in the *fourth* bit – the transmitted codeword is therefore 101101. Leaving out the check bits x_1, x_2 and x_4, the information message is therefore 101.

To obtain the codewords $x=x_1x_2x_3x_4x_5x_6$ we use the check equations $Hx'=0$, which by (2.25) become

$$x_4=x_5+x_6,\quad x_2=x_3+x_6,\quad x_1=x_3+x_5$$

Following the procedure described in Examples 2.19 and 2.20, by taking all possible combinations of the information bits x_3, x_5 and x_6, we can construct the whole set of codewords. For example, if $x_3=1$, $x_5=1$, $x_6=0$ then

$$x_1=1+1=0,\quad x_2=1+0=1,\quad x_4=1+0=1$$

and the codeword is 011110. Since each information bit can be independently 0

or 1, there are 2^3 possibilities, and we get the following table:

Check bits			Information bits			
x_1	x_2	x_4	x_3	x_5	x_6	Weight
0	0	0	0	0	0	0
1	1	0	1	0	0	3
1	0	1	0	1	0	3
0	1	1	0	0	1	4
0	1	1	1	1	0	4
1	0	1	1	0	1	4
1	1	0	0	1	1	4
0	0	0	1	1	1	3

The weight of each codeword is indicated in the final column. Since the smallest non-zero value is 3, Theorem 2.2 tells us that for this Hamming code with $r=3$, $n=6$ the minimum distance is $d=3$. Hence (by Theorem 2.1) this confirms that the code corrects all single errors.

Notice, however, that in this example if $m=000111$ then the syndrome is

$$s = Hm' = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \sim 111$$

which is the binary representation of 7 – but the code has length only 6, so there cannot be an error in the seventh bit! What this means is that more than one transmission error has occurred.

Let's now look at the properties of a general Hamming code having length n, dimension k and $r=n-k$ check bits (it is called an $[n, k]$ code).

(i) The $r \times n$ check matrix H has as its ith column the number i written in binary form, with $i=1, 2, \ldots, n$.

(ii) Since the columns of H satisfy the conditions of Theorem 2.4, it is the check matrix for an s.e.c. code.

(iii) The check bits are in the positions where the columns of H contain a single 1, that is positions $1, 2, 4, 8, \ldots, 2^{r-1}$.

(iv) For $n \geqslant 4$, the first few columns of H are

0 0 0 0 0 . . .
0 0 0 0 0 . . .
. . . . . .
. . . 1 1 . . .
. 1 1 0 0 . . .
1 0 1 0 1 . . .

Starting from the bottom row, these correspond to the check equations

$$x_1 = x_3 + x_5 + \cdots, \quad x_2 = x_3 + \cdots, \quad x_4 = x_5 + \cdots$$

When we construct the table listing the codewords, then the particular choice $x_3 = 1$, all other information bits zero, gives $x_1 = 1$, $x_2 = 1$, all other check bits zero.

Hence $111000\ldots0$ is always a codeword, with weight 3. We knew already that $d \geqslant 3$ since the code corrects all single errors, so we have now shown that Hamming codes have minimum distance exactly equal to 3.

(v) As we saw in the discussion following Theorem 2.4, the length n of any s.e.c. linear binary code is at most $2^r - 1$. When $n = 2^r - 1$ the Hamming code is called *perfect*, otherwise it is *shortened*.

(vi) For a perfect Hamming code, every syndrome (except $\mathbf{0}$) occurs as a single column of the check matrix, and so represents a single correctable error. For a shortened Hamming code, some syndromes represent multiple errors.

■ *EXAMPLE 2.26*

To construct the perfect [7, 4] Hamming code, we append onto the matrix H in (2.25) an extra column obtained from the binary representation 111 of 7. This produces the new check matrix

$$H = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix} \tag{2.26}$$

If a received message is $m = 1111001$, then the syndrome is

$$s = Hm' = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \sim 011$$

Since 011 is the binary representation of 3, we deduce that there is an error in the *third* bit, so the transmitted codeword is 1101001. Discarding the check bits in positions 1, 2, 4 leaves the decoded information message as 0001.

EXERCISE 2.27

(a) Give the check matrix for the Hamming code with six information bits and four check bits.

(b) Encode the information messages (i) 101101, (ii) 001011.

(c) Decode the received words (i) 0001110101, (ii) 0000101100, (iii) 1011110111, giving if possible the information messages which were sent.

EXERCISE 2.28 Starting with the check matrix in the previous exercise, write down the check matrix for the perfect Hamming code with four check bits. Decode the received word 101111011100000. How does this compare with (iii) in part (c) of the previous exercise?

An interesting geometrical representation can be given for the perfect Hamming code with three check bits, whose check matrix is set out in (2.26). This takes the form of a *Venn diagram*, shown in Figure 2.6. The information bits x_3, x_5, x_6, x_7 are located in the four central compartments of the diagram, indicated with hatched lines. The check bits x_1, x_2, x_4 are in the outer compartments. The check equations $Hx' = 0$ with H in (2.26) are

$$x_4 + x_5 + x_6 + x_7 = 0$$
$$x_2 + x_3 + x_6 + x_7 = 0$$
$$x_1 + x_3 + x_5 + x_7 = 0$$

However, the sum of the bits inside circle A in Figure 2.6 is precisely $x_4 + x_5 + x_6 + x_7$, which by the first check equation is zero; in other words, there must be an *even* number of ones inside circle A. You can easily confirm that the *same thing* holds for circles B and C, using the second and third check equations respectively. We can use this Venn diagram, in which each of the circles has even parity, both for encoding and decoding. For example, if the information message is $x_3x_5x_6x_7 = 1011$ then we obtain Figure 2.7(a). It is easy to see that the only way for each circle to contain an even number of ones must be as shown in Figure 2.7(b). Hence $x_1 = 0$, $x_2 = 1$, $x_4 = 0$ and the codeword is 0110011.

Suppose now that this codeword is transmitted, and an error occurs in the third bit. The received word 0100011 is shown in Figure 2.8. Both circles B and C contain an odd number of ones whereas A contains an even number of ones. We therefore deduce that the error lies at the intersection of circles B and C *only*. From Figure 2.6 we see that this intersection is x_3, so we conclude (correctly) that an error has occurred in the third bit.

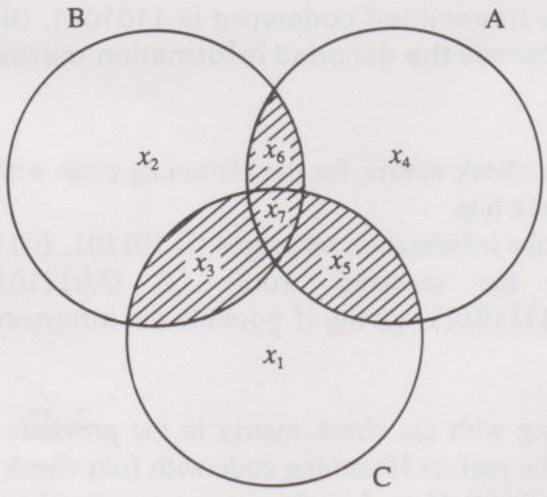


Figure 2.6

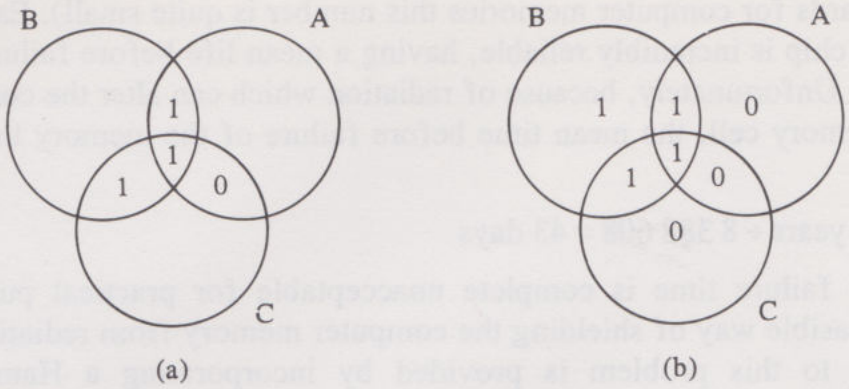


Figure 2.7

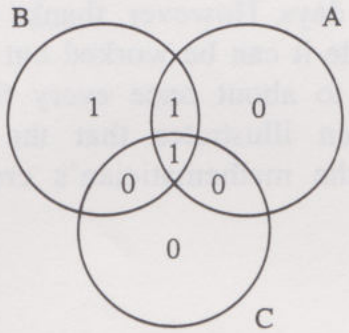


Figure 2.8

Notice, again by referring to Figure 2.6, that a single error in one of the check bits x_1, x_2 or x_4 is immediately recognized, since only one of the circles has its parity upset.

EXERCISE 2.29 Repeat the decoding problem in Example 2.26 using the Venn diagram representation.

EXERCISE 2.30 Use the Venn diagram representation to construct the set of codewords for the perfect [7, 4] Hamming code. Suppose that the codeword 0011001 is transmitted, and that errors occur in the second and sixth bits. Investigate what happens if the Venn diagram method is used to decode the received word.

An important application of Hamming codes is to improve dramatically the reliability of computer memories. Unavoidable errors occur in individual storage cells owing to stray radiation, which arises because the plastic packaging of memory chips contains small amounts of radioactive materials. For example, consider a memory bank consisting of 128 silicon 64K chips. Each chip contains $64 \times 2^{10} = 65\,536$ data storage cells, so that there is a total of $128 \times 65\,536 = 8\,388\,608$ such cells; in other words, the memory holds about a million 8 bit words (actually, by

today's standards for computer memories this number is quite small). Each memory cell in a 64K chip is incredibly reliable, having a mean life before failure of over a million years. Unfortunately, because of radiation which can alter the contents of an individual memory cell, the mean time before failure of the memory bank itself is about

$$1 \text{ million years} \div 8\,388\,608 \simeq 43 \text{ days}$$

Such a short failure time is complete unacceptable for practical purposes, but there is no feasible way of shielding the computer memory from radiation damage. The solution to this problem is provided by incorporating a Hamming code (actually a [64, 57] code with seven check bits) into the memory bank – this requires about 20% more capacity. This enlarged memory would be even more prone to errors (as there are more cells to be hit by radiation), the mean time to failure now being $43/1.2 \simeq 36$ days. However, thanks to the single-error correction provided by the Hamming code it can be worked out that the frequency of errors in the memory comes down to about once every 63 years (yes, years!). This remarkable improvement again illustrates that the wonders of contemporary electronics rely heavily on the mathematician's crucial contribution of error-correcting codes.

2.6 DECIMAL CODES

After concentrating on binary codes, it's time to take a closer look at decimal codes, which were introduced in Section 2.1. Let's return to the International Standard Book Number (ISBN), discussed in Example 2.8. Recall that this consists of a 10 digit codeword $x_1 x_2 \ldots x_{10}$ which uniquely defines a book. The last digit x_{10} is the check digit. This is chosen so that the *check sum*, defined by

$$S = \sum_{i=1}^{10} i x_i = x_1 + 2x_2 + 3x_3 + \cdots + 9x_9 + 10x_{10} \tag{2.27}$$

which is called the *weighted sum* of the digits, is a multiple of 11, that is

$$S = 0 \pmod{11}$$

The digits $x_1, x_2, \ldots, x_9$ can take any of the values $0, 1, 2, \ldots, 9$ but the check digit x_{10} can in addition take the value 10, which is denoted by the Roman numeral X. Setting the sum in (2.27) equal to zero shows that

$$-10x_{10} = x_1 + 2x_2 + \cdots + 9x_9$$

and because our arithmetic is modulo 11 then $-10x_{10} = x_{10}$ (since $11x_{10} = 0$). Hence for a given book number $x_1 \ldots x_9$ the check digit can be computed from

$$x_{10} = \sum_{i=1}^{9} i x_i \pmod{11} \tag{2.28}$$

A simple way of evaluating the check digit in (2.28) is suggested in Problem 2.8 at the end of the chapter.

In order to test whether a given book number is correct, that is a valid ISBN for a book, a simple tabular method was devised for use by librarians. The easiest way to understand this is to first consider a numerical example.

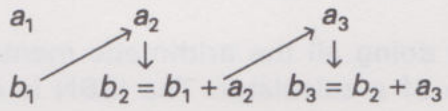
■ *EXAMPLE 2.27*

Consider the ISBN in Figure 2.2, namely 0-19-859665-0. Note that the hyphens are inserted purely for convenience in breaking up the book numbers into blocks, and have no mathematical significance. Write the book number in a vertical column denoted by c_1, and construct two columns c_2 and c_3 to the right as shown in Table 2.1.

The rules for constructing the table are as follows:

(i) The three entries in the first row are identical.

(ii) Suppose that at some stage the entries in a row are a_1, a_2, a_3. Then the row below it, b_1, b_2, b_3, is obtained from



The application of rule (ii) is indicated by arrows in Table 2.1.

The ISBN is correct if the last entry is a multiple of 11 – that is 0(mod 11) – so this final entry (circled in Table 2.1) represents an alternative check sum.

Table 2.1

c_1	c_2	c_3
0	0	0
1	1	1
9	10	11
8	18	29
5	23	52
9	32	84
6	38	122
6	44	166
5	49	215
0	49	(264) Check sum = 24 × 11

Notice that the effort involved in constructing Table 2.1 can be reduced by performing *each individual addition* modulo 11. For example, the fourth row of 8, 18, 29 becomes

$$8,\quad 8+10=7,\quad 7+11=7 \pmod{11}$$

and the next row is

$$5,\quad 5+7=1,\quad 1+7=8$$

You should check that the complete new table is as follows:

0	0	0
1	1	1
9	10	0
8	7	7
5	1	8
9	10	7
6	5	1
6	0	1
5	5	6
0	5	0

Clearly there is no difficulty in doing all the arithmetic mentally, unlike (2.27) which usually requires the use of a calculator. The ISBN is correct if the last entry in the table is zero.

We can now see what happens in general. Apply the rules (i) and (ii) to an arbitrary ISBN $x_1x_2 \ldots x_{10}$ to produce the following table:

c_1	c_2	c_3
x_1	x_1	x_1
x_2	$x_1 + x_2$	$2x_1 + x_2$
x_3	$x_1 + x_2 + x_3$	$3x_1 + 2x_2 + x_3$
x_4	$x_1 + x_2 + x_3 + x_4$	$4x_1 + 3x_2 + 2x_3 + x_4$
$\vdots$	$\vdots$	$\vdots$
x_{10}	$x_1 + x_2 + \cdots + x_{10}$	$T = (10x_1 + 9x_2 + 8x_3 + \cdots + 2x_9 + x_{10})$

The last entry T in column c_3 can be written as

$$T = 11(x_1 + x_2 + \cdots + x_9 + x_{10}) - S$$

where S is the check sum defined in (2.27). You can now see why T can be used as an alternative form of check sum, since clearly it is a multiple of 11 if and only if S is a multiple of 11.

EXERCISE 2.31 Test whether the following are correct ISBNs:

(a) 0-87-150702-2, (b) 0-13-152447-X.

EXERCISE 2.32 Use (2.28) to determine the check digit for an ISBN whose first nine digits are 039330711. A simplification of the procedure is given in Problem 2.8.

The ISBN code detects all single errors, for in this case the check sum S in (2.27) is not a multiple of 11. To see this, suppose that our proposed ISBN is $x_1 x_2 \ldots y_p \ldots x_{10}$, where the pth digit is $y_p = x_p + e$, where $e \neq 0$ is the error. Then the check sum S is

$$\begin{aligned} S &= x_1 + 2x_2 + 3x_3 + \cdots + py_p + \cdots + 10x_{10} \\ &= \sum_{i=1}^{10} ix_i + pe \\ &= pe(\text{mod } 11) \end{aligned} \tag{2.29}$$

since $x_1 \ldots x_{10}$ is a correct ISBN. The crucial fact that we now use is that since 11 is a *prime* number, there are no non-zero integers p and e such that their product is a multiple of 11. Hence S in (2.29) cannot be 0(mod 11), so the error has been detected. In fact, if it is known which digit is in error, then we can actually correct it. Again, this is best seen from a numerical example.

■ ***EXAMPLE 2.28***

Suppose that the fifth digit in the ISBN in Example 2.27 has been accidentally obliterated, so that it is recorded as 0198x96650. We wish to determine x. We construct the check sum, either using the tabular method, or directly from (2.27). The latter gives

$$\begin{aligned} S &= 1.0 + 2.1 + 3.9 + 4.8 + 5.x + 6.9 + 7.6 + 8.6 + 9.5 + 10.0 \\ &= 250 + 5x \\ &= 8 + 5x(\text{mod } 11) \end{aligned}$$

and the required value of x is that which makes S a multiple of 11. Simply by trying successive values of $x = 1, 2, 3, \ldots$ we find that $x = 5$ gives $S = 33 = 0(\text{mod } 11)$.

EXERCISE 2.33 An ISBN is received with one digit illegible, namely 09481□4786. Determine the missing digit.

If a received ISBN is found to be incorrect and it is not known which digits are in error, then there will be many possible correct ISBNs. Even if we make our usual assumption that a single error is the most likely occurrence we will not be able to correct it.

EXERCISE 2.34 An ISBN is received as 0-471-62187-3.

(a) Show that it is incorrect.
(b) Determine the correct check digit, assuming the other digits are correct.
(c) Assume that the original check digit is correct, but there is a single error in one of the digits $x_2, x_3, \ldots, x_9$. Determine three possible ISBNs.

It should be stressed that in Example 2.28 and in problems like Exercise 2.33 a *unique* answer is always obtained. This is because the set of integers modulo 11 which we use for the ISBN is an example of what is called a *finite field*, or *Galois field*, after the French mathematician who was killed in a duel in 1832 at the age of only 20 (he was fighting not about mathematics, but over a matter of 'honour'!). Without becoming technical, for our purposes we can think of a finite field (of *order* N) as a set of N elements which it is possible to add, subtract, multiply and divide (but division by 0 is not defined). The notation $GF(N)$ is commonly used. Every element α in the field has an additive inverse β, which means that there is another element β such that $\alpha+\beta=0$. Every non-zero element α has a multiplicative inverse, which means that there is an element γ in the field such that $\alpha\gamma=1$. We write $-\alpha$ for β, and α^{-1} for γ.

■ EXAMPLE 2.29

We mentioned in Section 2.1 that modular arithmetic provides a way of constructing finite fields. Consider the set of integers 0, 1, 2, ..., 10 where we perform arithmetic modulo 11. This set of 11 numbers forms $GF(11)$. Addition and subtraction are as described earlier. The additive inverse is easily handled, for example

$$9+2=0 \pmod{11}, \quad \text{so } -9=2$$

The multiplicative inverse requires a little more thought. For example, to find this for the integer 9, simply evaluate

$$9\times 1=9, \quad 9\times 2=18=7 \pmod{11}, \quad 9\times 3=27=5 \pmod{11}, \quad \ldots$$

until we find a product which is $1 \pmod{11}$. Clearly $9\times 5=45=1 \pmod{11}$, so we have found that $9^{-1}=5$.

EXERCISE 2.35 For each of the remaining 10 members of the set of integers modulo 11, determine the additive and multiplicative inverse.

EXERCISE 2.36 For any non-zero element a in a finite field F there exists a multiplicative inverse a^{-1} such that $a^{-1}a=1$. Use this to prove that if a, b are any elements in F then $ab=0$ implies *either* $a=0$ *or* $b=0$ (you may assume that the usual distributive, associative and commutative laws apply for finite fields, namely $a+b=b+a$, $ab=ba$, $(a+b)+c=a+(b+c)$, $(ab)c=a(bc)$ and $a(b+c)=ab+ac$). Note that this result would *not* apply, for example, to the set of integers modulo 10, since $5\times 2=0 \pmod{10}$.

In fact, finite fields can consist of sets of polynomials, as well as of integers, but exploring this subject any further is beyond the scope of this book. It is most interesting to realize, however, that what was once thought to be a purely abstract piece of mathematics is now a crucial tool in the development of error-correcting codes.

In transmitting or recording information in digital form, a common error is to interchange (or 'transpose') accidentally two digits, usually neighbouring ones. We now show that the ISBN can detect any double error of this type. Suppose that $x_1 \ldots x_{10}$ is a correct ISBN but the received number is

$$x_1 x_2 \ldots x_k \ldots x_j \ldots x_{10}$$

x_j and x_k interchanged in error

The check sum (2.27) now becomes

$$S = 1x_1 + 2x_2 + \cdots + jx_k + \cdots + kx_j + \cdots + 10x_{10} \tag{2.30}$$

In order to relate this sum to (2.27), add and subtract terms kx_k, jx_j to (2.30) as follows:

$$\begin{aligned} S &= 1x_1 + 2x_2 + \cdots + jx_k - kx_k + kx_k + \cdots + kx_j - jx_j + jx_j + \cdots + 10x_{10} \\ &= 1x_1 + 2x_2 + \cdots + (j-k)x_k + kx_k + \cdots + (k-j)x_j + jx_j + \cdots + 10x_{10} \\ &= \sum_{i=1}^{10} ix_i + (j-k)x_k + (k-j)x_j \end{aligned} \tag{2.31}$$

$$\delta = (j-k)(x_k - x_j) \pmod{11} \tag{2.32}$$

where the first term in (2.31) is $0 \pmod{11}$ because $x_1 x_2 \ldots x_{10}$ is a correct ISBN. The expression in (2.32) cannot be $0 \pmod{11}$, because $j \neq k$, and $x_j \neq x_k$ (if $x_j = x_k$ there is no error!), and we have seen in Exercise 2.36 that a key property of finite fields is that the product of any two non-zero elements is non-zero. Therefore, the check sum (2.30) is not a multiple of 11, so errors involving the interchange of *any* two unequal digits are always detected.

EXERCISE 2.37 Verify that if the ISBN in Example 2.27 is incorrectly recorded as 0-19-895665-0 then the transposition error is detected.

We now introduce a decimal code which, unlike the ISBN, will *correct* all single errors. This consists of all the codewords $x_1 x_2 \ldots x_{10}$ which satisfy the *two* check equations

$$S_1 = \sum_{i=1}^{10} x_i = 0 \pmod{11} \tag{2.33}$$

$$S_2 = \sum_{i=1}^{10} ix_i = 0 \pmod{11} \tag{2.34}$$

where the x_i take the values 0, 1, 2, ..., 8, 9, there now being *two* check digits x_9 and x_{10}. Notice that S_2 is the same as S in (2.27), so the second check equation above is the same as the single check equation for the ISBN, but now we do not allow x_{10} to take the value X (= 10). Suppose that a single error of magnitude e ($\neq 0$) occurs in the jth position, so that the received word is

$$x_1 x_2 \ldots x_{j-1}, \quad x_j + e, \quad x_{j+1} \ldots x_{10}$$

From the definition (2.33) the first check sum for this word is

$$S_1 = \sum_{i=1}^{10} x_i + e = e(\text{mod } 11)$$

showing that the magnitude of the error is equal to S_1(mod 11). Similarly, from (2.34) the second check sum is

$$S_2 = \sum_{i=1}^{10} ix_i + je = je(\text{mod } 11)$$

Bearing in mind throughout that we are working with the finite field $GF(11)$, we can write

$$j = S_2 e^{-1} = S_2 S_1^{-1}$$

which gives the position of the error (reading from left to right in the received word).

■ EXAMPLE 2.30

Suppose that a received word is 1764753052. To obtain S_1 in (2.33), sum the digits to get $S_1 = 40 = 7$, which is the magnitude e of the error. The weighted sum $S_2 = x_1 + 2x_2 + \cdots + 10x_{10}$ in (2.34) is easily found to be $S_2 = 200 = 2$. Hence the position of the error is $j = 2.7^{-1} = 2.8 = 16 = 5$ (you were asked to find multiplicative inverses (mod 11) in Exercise 2.35; here $8.7 = 56 = 1$ so $7^{-1} = 8$). The correct codeword is therefore 1764053052 (notice that the corrected fifth digit is obtained by *subtracting* the error e from the incorrect received fifth digit).

In general, the *decoding scheme* is as follows:

(i) If $S_1 = 0$, $S_2 = 0$ then assume there is no error and the received word is a codeword.

(ii) If $S_1 \neq 0$, $S_2 \neq 0$ then we assume that a single error has occurred in the digit with position $S_2 S^{-1}$, and this digit is corrected by subtracting S_1 from it.

(iii) If $S_1 = 0$ or $S_2 = 0$ (not both) then we have detected at least two errors.

It is interesting to realize that the above 10 digit code could be used as a telephone number code. There are over 82 million numbers which satisfy the two check equations (2.33) and (2.34), enough for all the telephones in the UK. The telephone exchange could use the decoding scheme described, so that if you keyed in one wrong digit you would still have your call placed correctly; and (see Exercise 2.39 below) if you inadvertently interchanged two digits (easily done!) then instead of getting a 'wrong number' you would get an error signal ('number unobtainable' tone). Just think how much frustration would be saved if a code like this was adopted countrywide!

EXERCISE 2.38 Decode the received words 0206211909 and 061293587 using the above code.

EXERCISE 2.39 Show that the above code detects all errors where two (different) digits of a codeword have been interchanged.

EXERCISE 2.40 Show that the decimal code consisting of all words $x_1x_2 \ldots x_{10}$, with $x_i \in \{0, 1, 2, \ldots, 9\}$, satisfying the check equations

$$\sum_{i=1}^{10} x_i = 0 (\text{mod } 10), \quad \sum_{i=1}^{10} i x_i = 0 (\text{mod } 10)$$

is *not* a single-error-correcting code. (Hint: find two codewords which are distance 2 apart.)

We conclude this chapter with a decimal code of length 10, defined as before over $GF(11)$, but which corrects all *double* errors. We begin with the s.e.c. decimal code just described, and select those codewords which satisfy in addition to (2.33) and (2.34) the extra two check equations

$$S_3 = \sum_{i=1}^{10} i^2 x_i = 0, \quad S_4 = \sum_{i=1}^{10} i^3 x_i = 0$$

Suppose that two errors of magnitudes e_1 and e_2 occur in positions i and j respectively. Using the same argument as for the s.e.c. case we obtain

$$S_1 = \sum_{i=1}^{10} (x_i + e_1 + e_2) = e_1 + e_2 \tag{1}$$

$$S_2 = ie_1 + je_2 \tag{2}$$

$$S_3 = i^2 e_1 + j^2 e_2 \tag{3}$$

$$S_4 = i^3 e_1 + j^3 e_2 \tag{4}$$

where all arithmetic is modulo 11. We therefore have the four equations (1) to (4) to solve for the four unknowns i, j, e_1, e_2. Although these equations are non-linear in i

and j, they can be put into a form which is much easier to solve. The trick is to eliminate e_1, e_2 and i so as to obtain a quadratic equation in j. Carry out the operations indicated below, for example, the first one is to subtract equation (2) from j times equation (1).

$$j\times(1)-(2): \quad jS_1 - S_2 = (j-i)e_1 \tag{5}$$

$$j\times(2)-(3): \quad jS_2 - S_3 = (j-i)ie_1 \tag{6}$$

$$j\times(3)-(4): \quad jS_3 - S_4 = (j-i)i^2e_1 \tag{7}$$

$$(6)^2-(5)\times(7): \quad (jS_2 - S_3)^2 = (jS_1 - S_2)(jS_3 - S_4)$$

Rearranging the last equation gives

$$j^2(S_2^2 - S_1S_3) + j(S_1S_4 - S_2S_3) + (S_3^2 - S_2S_4) = 0 \tag{2.35}$$

EXERCISE 2.41 Verify that by eliminating e_1, e_2 and j from the equations (1) to (4) you obtain a quadratic equation in i with the *same* coefficients as those in (2.35).

In view of the above exercise, we conclude that the locations of the errors are the two roots of the quadratic equation

$$ax^2 + bx + c = 0 \tag{2.36}$$

where

$$a = S_2^2 - S_1S_3, \quad b = S_1S_4 - S_2S_3, \quad c = S_3^2 - S_2S_4 \tag{2.37}$$

Once i and j have been found, it is very easy to determine e_1 and e_2 from the two linear equations (1) and (2). Notice that if just one error occurs (say $e_1 \neq 0$, $e_2 = 0$) then

$$S_1 = e_1, \quad S_2 = ie_1, \quad S_3 = i^2e_1, \quad S_4 = i^3e_1$$

in which case substituting into (2.37) gives $a=0$, $b=0$, $c=0$.

We can summarize the decoding scheme as follows:

(i) If $S_1 = S_2 = S_3 = S_4 = 0$, then assume there is no error and the received word is a codeword.

(ii) If $\mathbf{S} = (S_1, S_2, S_3, S_4) \neq \mathbf{0}$ and $a = b = c = 0$ then we assume that a single error has occurred in digit position $S_2S_1^{-1}$, which is corrected by subtracting S_1 from it (just as for the s.e.c. code).

(iii) If $\mathbf{S} \neq \mathbf{0}$ and $a \neq 0$, $b \neq 0$ then (2.36) has solutions

$$i, j = \frac{-b \pm \sqrt{(b^2 - 4ac)}}{2a} \tag{2.38}$$

provided $b^2 - 4ac$ is a non-zero square in $GF(11)$. We assume there are two errors in positions i, j. These digits are corrected by subtracting

respectively e_1 and e_2 from them, where

$$e_2 = \frac{iS_i - S_2}{i - j}, \quad e_1 = S_1 - e_2 \tag{2.39}$$

(iv) If none of (i), (ii) or (iii) holds, then we have detected at least three errors.

An explanation is necessary of what we mean by square roots in $GF(11)$. For example,

$$7^2 = 49 = 5 \pmod{11}$$

so we can write $\sqrt{5} = 7$. Construct the following table of squares of integers less than 11:

x	1	2	3	4	5	6	7	8	9	10
x^2	1	4	9	5	3	3	5	9	4	1

If we read from the bottom row to the top this gives all the available square roots. This can be written as follows:

y	1	3	4	5	9
$\sqrt{y}$	1	5	2	4	3

Notice that in some cases the square root is not unique. For example, we see that $4^2 = 7^2 = 5$, so that $\sqrt{5} = 4$ or 7. However, this does not affect the solutions of (2.36), because $4 = -7 \pmod{11}$ and therefore $\pm 4 = \mp 7$ in (2.38), showing that the same pair of values i, j is obtained whichever square root is used.

Since 2, 6, 7, 8 do not have square roots (mod 11) (they do not occur in the bottom row of the table of squares) then if $b^2 - 4ac$ in (2.38) takes one of these values we have detected more than two errors.

Using techniques beyond the scope of this book it can be shown that the minimum distance for this code is 5, verifying by Theorem 2.1 that the code does indeed correct all double errors.

■ *EXAMPLE 2.31*

For the above code, decode the received word 1204000910.

We compute, using modulo 11 arithmetic throughout:

$$S_1 = \sum x_i = 1 + 2 + 4 + 9 + 1 = 17 = 6$$

$$S_2 = \sum ix_i = 1.1 + 2.2 + 4.4 + 8.9 + 9.1 = 102 = 3$$

$$S_3 = \sum i^2 x_i = 1^2.1 + 2^2.2 + 4^2.4 + 8^2.9 + 9^2.1 = 730 = 4$$

$$S_4 = \sum i^3 x_i = 1^3.1 + 2^3.2 + 4^3.4 + 8^3.9 + 9^3.1 = 5610 = 0$$

From (2.37) we have

$$a = S_2^2 - S_1 S_3 = 9 - 24 = -15 = 7$$
$$b = S_1 S_4 - S_2 S_3 = 0 - 12 = 10$$
$$c = S_3^2 - S_2 S_4 = 16 - 0 = 5$$

so that case (iii) of the decoding scheme applies. We therefore need to solve the quadratic equation (2.36), namely

$$7x^2 + 10x + 5 = 0$$

which has roots

$$i, j = \frac{-10 \pm \sqrt{(100 - 140)}}{14}$$
$$= \frac{-10 \pm \sqrt{(-40)}}{3}$$
$$= \frac{-10 \pm \sqrt{4}}{3} = \frac{-10 \pm 2}{3}$$
$$= -8 \times 3^{-1}, \quad -12 \times 3^{-1}$$

These roots reduce as follows:

$$-8 \times 3^{-1} = 3 \times 4 = 12 = 1$$
$$-12 \times 3^{-1} = 10 \times 4 = 40 = 7$$

Hence the errors occur in digits $i = 1$ and $j = 7$. From (2.39)

$$e_2 = \frac{1.6 - 3}{1 - 7} = 3(-6)^{-1} = 3.5^{-1} = 3.9 = 27 = 5$$
$$e_1 = 6 - 5 = 1$$

so the corrected first and seventh digits are respectively $1 - 1 = 0$, $0 - 5 = 6$, and the decoded word is 0204006910.

EXERCISE 2.42 For the above decimal code, decode the received word 4003100711.

The decimal code described above is an example of an important class of codes called BCH codes, which can be developed to correct more than two errors – for example, in long-distance telecommunications codes of length 255 with 24 check bits are used to correct three errors. Details of these codes belong in more advanced books.

PROBLEMS

2.1 The United States 'Zip' postcode was introduced in Example 2.6, with a particular case reproduced in Figure 2.4. The code is a number $a_1 a_2 \ldots a_9$ with nine decimal digits. The digits are represented in a machine-readable barcode form, according to

the following scheme, where 0 corresponds to a short bar and 1 to a long bar:

Decimal digit	Barcode
1	00011
2	00101
3	00110
4	01001
5	01010
6	01100
7	10001
8	10010
9	10100
0	11000

Each digit is represented by a block containing two long and three short bars, for example 8 is |||||. Notice that there are no other possible arrangements of two long and three short bars.

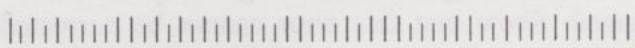
Ignoring the long 'spacer' bars at each end of the code, we see in Figure 2.4 that there are precisely 50 bars, which represent a 10 digit codeword $a_1 a_2 \ldots a_{10}$. The extra digit a_{10} is the check digit defined by

$$\sum_{i=1}^{10} a_i = 0 (\text{mod } 10)$$

That is, the sum of the digits is a multiple of 10.

(a) The Zip code for North Carolina State University is 27695-8205. Determine the check digit.

(b) Determine the Zip code represented by the barcode



(c) Show that if the machine makes a single error in reading a barcode (i.e. a short bar is read as a long bar, or vice versa) then this error is always detected.

(d) Show that if there is a single error, then once the location of the block in which this occurs has been determined, the error can be corrected.

Hence determine the correct Zip code if the following barcode contains a single error.

(e) Explain why if any *two* errors occur in a particular block of five bars, then these can always be detected. Explain also why some, but not all, such double errors can be corrected.

2.2 The United States Postal Service money-order identification number, introduced in Example 2.9, consists of 10 decimal digits and a check digit. The check digit is the

remainder modulo 9 of the 10 digit number, that is $x_1x_2 \ldots x_{10} = x_{11} \pmod 9$.

(a) Prove that a correct identification number satisfies

$$x_1 + x_2 + \cdots + x_{10} = x_{11} \pmod 9$$

(b) Deduce that errors in which a 9 is replaced by a 0, or vice versa (excluding the check digit), go undetected.

(c) Show also that errors involving the transposition of two adjacent digits will only be detected if they involve the check digit. What happens if the two transposed digits are not adjacent?

2.3 A codeword $x_1x_2x_3x_4x_5$ consisting of five decimal digits is defined so that the check digit x_5 satisfies the condition

$$x_1 + 3x_2 + 7x_3 + 9x_4 + x_5 = 0 \pmod{10}$$

(a) Encode the number 9347.

(b) Show that all single errors will be detected.

(c) Will all transpositions of adjacent digits be detected?

(d) Show that the 'weights' 1, 3, 7, 9 multiplying x_1, x_2, x_3, x_4 respectively are the only possible integers less than 10 which can be used if all single errors are to be detected.

2.4 Write down the check matrix for the perfect Hamming code with four check bits. Encode the information message 10110111101. If a received word is 010101100011010, determine the information bits of the transmitted message.

2.5 An International Standard Book Number (ISBN) is recorded as 0-19-853827-3. Show that it is incorrect. Determine the correct ISBN, assuming that the sixth digit (reading from left to right) is in error.

It is later found that the first six digits *were* correctly recorded, but that two other adjacent digits in the six-digit number assigned by the publisher had been inadvertently interchanged. Show that this error cannot be corrected, by determining two possible ISBNs which satisfy the given conditions.

2.6 Since 1966 all Norwegian citizens have been allocated an identification number $x_1x_2x_3 \ldots x_{11}$ consisting of 11 decimal digits. The first six digits represent the date of birth (day, month, year), $x_7x_8x_9$ is a personal number and x_{10}, x_{11} are check digits defined by

$$x_{10} = -(3x_1 + 7x_2 + 6x_3 + x_4 + 8x_5 + 9x_6 + 4x_7 + 5x_8 + 2x_9) \pmod{11}$$

$$x_{11} = -(5x_1 + 4x_2 + 3x_3 + 2x_4 + 7x_5 + 6x_6 + 5x_7 + 4x_8 + 3x_9 + 2x_{10}) \pmod{11}$$

Write down the check matrix for this code. Hence verify that the code will not detect double errors of the form $x_4 \pm \varepsilon$, $x_{10} + 11 \mp \varepsilon$, where ε can take any of the values $0, 1, 2, \ldots, 10$.

2.7 The double-error-correcting decimal code described in Section 2.6 can be extended to correct more than two errors. For example, to correct three errors we use *six* check

equations $S_1 = 0, S_2 = 0, \ldots, S_6 = 0 \pmod{11}$, where

$$S_{j+1} = \sum_{i=1}^{10} i^j x_i, \quad j = 0, 1, 2, 3, 4, 5$$

The positions p_1, p_2, p_3 of the errors are then given by the three roots of the cubic equation

$$x^3 + a_1 x^2 + a_2 x + a_3 = 0$$

where the a_i are the solutions of the equations

$$S_4 + a_1 S_3 + a_2 S_2 + a_3 S_1 = 0$$
$$S_5 + a_1 S_4 + a_2 S_3 + a_3 S_2 = 0$$
$$S_6 + a_1 S_5 + a_2 S_4 + a_3 S_3 = 0$$

Since all arithmetic is performed modulo 11, the roots of the cubic equation are found simply by trying the values $x = 0, 1, 2, \ldots, 10$.

The corresponding magnitudes e_1, e_2, e_3 of the errors are then given by the solution of the set of equations

$$\begin{bmatrix} 1 & 1 & 1 \\ p_1 & p_2 & p_3 \\ p_1^2 & p_2^2 & p_3^2 \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix} = \begin{bmatrix} S_1 \\ S_2 \\ S_3 \end{bmatrix}$$

Suppose that a word is received for which it is found that

$$S_1 = 2, \quad S_2 = 8, \quad S_3 = 4, \quad S_4 = 5, \quad S_5 = 3, \quad S_6 = 2$$

Assuming that three transmission errors have occurred, find their positions and magnitudes.

2.8 A simple way of determining the check digit x_{10} for an ISBN using the expression (2.28) is to apply the fact that the remainder modulo 11 of a three-digit decimal number abc is equal to $(c - b + a) \pmod{11}$. For example, if the first nine digits of an ISBN are 039330711 (see Exercise 2.32) then from (2.28) we have

$$\begin{aligned} x_{10} &= \sum_{i=1}^{9} i x_i \pmod{11} \\ &= 126 \pmod{11} \\ &= (6 - 2 + 1) \pmod{11} = 5 \end{aligned}$$

Prove the stated fact. Also prove that in general for any n-digit decimal number $a_1 a_2 \ldots a_n$, its remainder (mod 11) is equal to

$$(a_n - a_{n-1} + a_{n-2} - a_{n-3} + \cdots) \bmod 11$$

2.9 The Universal Product Code used in the United States to identify retail products consists of a number $x_1 x_2 x_3 \ldots x_{11} x_{12}$ with 12 decimal digits. The first digit denotes the product type, the next five digits represent the manufacturer, and the next five

digits are assigned to the product by the manufacturer. The final digit x_{12} is the check digit defined by

$$3(x_1 + x_3 + x_5 + x_7 + x_9 + x_{11}) + x_2 + x_4 + x_6 + x_8 + x_{10} + x_{12} = 0 (\text{mod } 10)$$

(a) Prove that the code detects all single errors.
(b) Does the code detect all errors involving the transposition of two adjacent digits?

2.10 The Article Number Association (ANA) System in the UK is a similar scheme to that in the previous problem. It consists of a number with 13 decimal digits, where the check digit x_{13} is defined by

$$3(x_2 + x_4 + x_6 + x_8 + x_{10} + x_{12}) + x_1 + x_3 + x_5 + x_7 + x_9 + x_{11} + x_{13} = 0 (\text{mod } 10)$$

and is based on the EAN system described in Example 2.5. For example, the *Guardian* newspaper on Mondays has the number 9770261307019. Here $x_1x_2x_3$ denotes the product type, x_4 retrieves the price from the memory of the shop's computer, the next seven digits are the *Guardian*'s code and x_{12} is the day of the week, starting with 1 for Monday, 2 for Tuesday and so on. What is the *Guardian*'s number on Fridays?

2.11 (a) A simple decimal code consists of words $x_1x_2 \ldots x_nc$, where the check digit c is defined by

$$x_1 + x_2 + \cdots + x_n = c (\text{mod } 10)$$

Deduce that the code detects all single errors in the digits x_i, but that no transpositions of adjacent x digits are detected.
(b) An improvement is obtained by using the check equation

$$x_1 - x_2 + x_3 - x_4 + x_5 - \cdots = c (\text{mod } 10)$$

Show that this still detects all single errors in the x digits, and also detects any error in which x_i and x_{i+1} are interchanged except when $x_i - x_{i+1} = \pm 5$. Hence deduce that 8/9 of all such transpositions are detected.

FURTHER READING

BACKHOUSE, J.K. 1983. 'Retail article numbering and bar codes', *IMA Bulletin*, **19**, 17–18.
BERNARD, J. 1986. 'Compact discs bit-by-bit', *Radio Electron.*, August, 62–3, 85.
BERRY, J., BURGHES, D. and HUNTLEY, I. 1986. *Decision Mathematics*. Ellis Horwood, Chichester, Chapter 13.
BRINN, L.W. 1984. 'Algebraic coding theory in the undergraduate curriculum', *Am. Math. Mon.*, **91**, 510–12.
COMAP. 1994. *For All Practical Purposes*. Freeman, New York, Chapter 9.
CONNOR, S. 1984. 'The invisible border guard', *New Sci.*, 5 January, 9–14.
GALLIAN, J.A. 1986. 'The Zip code bar code', *UMAP J.*, 7, 191–5.
GALLIAN, J.A. and WINTERS, S. 1988. 'Modular arithmetic in the marketplace', *Am. Math. Mon.*, **95**, 548–51.

HILL, R. 1986. *A First Course in Coding Theory*. Oxford University Press, Oxford.

HILL, R. 1989. 'Error-correcting codes I', *Math. Spectrum*, **22**(3), 94–103.

HILL, R. 1990. 'Error-correcting codes II', *Math. Spectrum*, **23**(1), 14–23.

McELIECE, R.J. 1985. 'The reliability of computer memories', *Sci. Am.*, January, 68–73.

SAVIR, D. and LAURER, G.J. 1975. 'The characteristics and decodability of the Universal Product Code symbol', *IBM Syst. J.*, **14**(1), 16–34.

SELMER, E.S. 1967. 'Registration numbers in Norway: Some applied number theory and psychology', *J. R. Stat. Soc. (Ser. A)*, **130**, 225–31.

TUCHINSKY, P.M. 1985. 'International Standard Book Numbers', *UMAP J.*, **5**, 41–54.

3

Making Things Happen

3.1 INTRODUCTION AND EXAMPLES

Perhaps one of the basic human drives is the desire to have control over events. From ancient times people have appealed to their deity to 'make things happen' as they would like them to. As technology has developed, devices have been invented which perform a controlling function without the need for human supervision. One of the oldest of these mechanisms can be traced back to the Greek civilization of the first century AD: you will certainly be familiar with the domestic toilet where, after flushing, the water tank automatically fills up to its original level. Indeed, if you lift off the cistern lid you are likely to find a line inside marking the normal water level. The job of the water-supplying device is to maintain the water in the tank as close as possible to this fixed level. Identical mechanisms operate in many households for the main cold water cistern; and for the 'header' tank which ensures that the radiators of a hot water central heating system are kept full of water. The principle behind the water regulating system is simple but ingenious, and is illustrated in Figure 3.1.

A floating ball monitors the water level; when this is below the desired level the supply valve is open and water flows into the cistern. The float rises with the water level, and when this reaches its preset position the supply valve shuts off. The valve is a purely mechanical device which is operated by the rigid arm attached to the float. This device for regulating the water level in a cistern exhibits several characteristics which

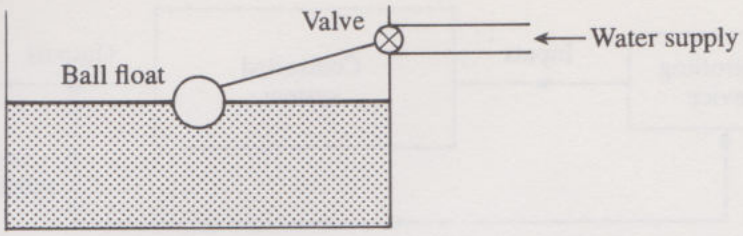


Figure 3.1

are common to many control systems. First, the operation is *automatic* – that is, when working correctly it needs no human intervention; secondly, it uses the idea of *feedback* – which means that a knowledge of the actual water level is 'fed back' by means of the float and arm so as to control the supply valve; thirdly, it will still work well even if there are wide fluctuations in the pressure of the incoming water – for example, if the water pressure drops the tank will simply take longer to be filled up.

You may have noticed the introduction of the expression *control system*: this is a term which is used in a wide sense to describe many situations. Some examples of control systems are:

(i) An aircraft, where one of many control problems is to fly and land using the 'autopilot'.

(ii) A manufacturing process, where the objective is to control the cost and quality of the end product.

(iii) The human body, where many functions are regulated automatically without our conscious intervention; for example, body temperature is kept so nearly constant that if it rises by even 1 degree we suspect that we are ill; when playing tennis or other games we are able to keep our eyes on the ball even though we are moving around vigorously.

(iv) A motor vehicle, which nowadays is full of control devices – for example, the fuel injection system which controls the fuel supply to the engine; an automatic gearbox; an antilock braking system which prevents the car skidding when the brakes are applied suddenly.

(v) The economy of a country or region where it is required to control, for example, the level of unemployment and the rate of inflation.

You should be aware that although the control systems mentioned above are all quite different from each other, they do have certain features in common. We can think of them as consisting of a lot of interconnected parts whose interactions we may or may not fully understand. Our aim is to make the system behave in some desired fashion by suitably controlling the *inputs* (or *control variables*) so as to produce satisfactory *outputs*. For example, in the plumbing system of Figure 3.1, the output is the water level, and the input is the water supply to the cistern. Don't imagine, however, that it's always necessary to have an accurate mathematical model of a system before it can be controlled. After all, most people manage to learn the tricky balancing act of riding a bicycle without ever thinking of the mathematics involved!

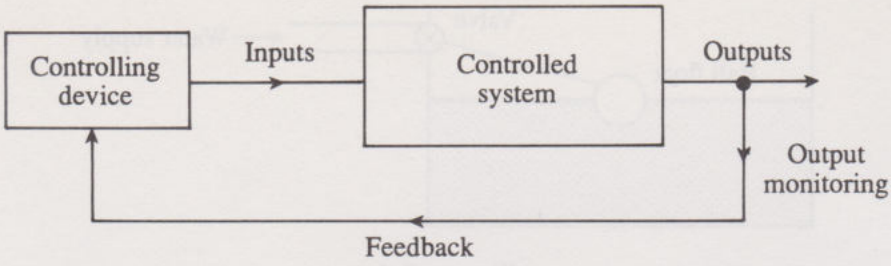


Figure 3.2

The idea of *feedback* is a crucial one. To take another homely example: in a central heating system the room thermostat is set to a comfortable level (say 20 °C). The thermostat monitors the actual temperature and 'feeds back' this information to the heat supply, turning it on or off according to whether the room temperature is above or below the desired setting. The input here is the heat supplied by the boiler or other heat source, and the output is the room temperature.

We can represent a typical situation as in Figure 3.2. The outputs are measured and this information is fed back to the controlling device which modifies the inputs (if necessary) so as to produce the desired behaviour. For example, an automatic landing system for an aircraft will be designed to ensure that at the instant when the wheels touch the runway, the aircraft has zero velocity in the vertical direction, so that it touches down without a bump. In order to achieve this many items such as altitude, airspeed, position of flaps and rudder will have to be monitored, and this information is fed back to the autopilot which is flying the aircraft. Because of the appearance of the diagram in Figure 3.2 the complete set-up is called a *closed loop system*, since the information from the output flows back around the 'loop'. In real life there will often be unpredictable external disturbances which affect the system under consideration — for example, sudden gusts of wind can occur as an aircraft approaches the runway, or the temperature outside a centrally heated building may suddenly drop. A well-designed control system will be able to cope with such unexpected influences.

During the nineteenth century many feedback devices were invented, the most famous and widely used being James Watt's governor for regulating the speed of steam engines. An early attempt to investigate some mathematical problems of control was made by Airy in 1840 when he was Astronomer Royal. He was interested in keeping a telescope pointed at a fixed point in the sky even though the Earth is rotating. The book by Mayr (1970) gives a good account of the origins of feedback control.

Let's now look at some examples of mathematical models of control systems.

■ EXAMPLE 3.1

Let's return to the bank savings model introduced in Example 1.1 in Chapter 1. Recall that $x(k)$ denotes the amount in the account at the end of the kth time

period, and that at the end of each period an amount of interest is added at the rate of $r/100n$ on the balance at the beginning of the period. A net amount of $u(k)$ is deposited in the account during the kth period, but this does not earn interest until the next time period (a negative $u(k)$ indicates a net withdrawal). The equation describing the behaviour of the account was found to be

$$x(k+1) = \left(1 + \frac{r}{100n}\right)x(k) + u(k+1), \quad k = 0, 1, 2, \ldots \tag{3.1}$$

Here the variable $x(k)$ denotes the *state* of the account at the end of the kth period, and in this example the output is simply equal to $x(k)$. The *control variable* $u(k)$ is used to determine a desired pattern of savings; that is, we exercise control over the sequence of outputs $x(1)$, $x(2)$, $x(3)$, ... by selecting a suitable sequence of inputs $u(1)$, $u(2)$, $u(3)$, ...

A more general version of (3.1) is

$$x(k+1) = \alpha x(k) + \beta u(k), \quad k = 0, 1, 2, \ldots \tag{3.2}$$

where α and β are constants. We have relabelled the sequence of inputs as $u(0)$, $u(1)$, $u(2)$, ... in (3.2) instead of $u(1)$, $u(2)$, $u(3)$, ··· in (3.1), so that we have $u(k)$ instead of $u(k+1)$ on the right-hand side of the equation. This is purely a convention, but one which is widely used.

EXERCISE 3.1 Consider the savings account described in Exercise 1.2 in Chapter 1. You were asked to show that, with the given sequence of inputs, the balance in the account at the end of 2 years would be £2493.96. Find how much you should deposit into the account so that after another 6 months have elapsed the account holds exactly £3000.

An important area of control applications involves dynamical systems which obey *Newton's law of motion*. This says that a body of mass m upon which a force u is exerted experiences an acceleration f where

$$mf = u \tag{3.3}$$

■ *EXAMPLE 3.2*

Suppose a car having mass m is being driven along a straight, level road. For simplicity assume that the car is controlled only by the throttle, producing an accelerating force u_1 on the car, and by the brake which produces a retarding force u_2 upon the car. Both of these forces will vary with time t in a continuous fashion, so we write $u_1(t)$, $u_2(t)$ to express the fact that u_1 and u_2 are *functions* of time. Suppose that we are only interested in the car's distance x_1 from some starting point, and its velocity x_2. Again, these quantities $x_1(t)$ and $x_2(t)$ will depend upon the time t which has elapsed since starting off. The acceleration f will be the *derivative* of the velocity, that is

$$f = \frac{dx_2}{dt}$$

so Newton's equation (3.3) becomes

$$m\frac{dx_2}{dt} = u_1 - u_2 \tag{3.4}$$

Notice that the total forces on the car are $u_1 - u_2$; the force u_2 carries a negative sign since it opposes the motion – that is, it has the effect of reducing the velocity. To complete the picture we need to add the fact that

$$\frac{dx_1}{dt} = x_2 \tag{3.5}$$

which states that the velocity at any instant is the derivative of the distance travelled. The two equations (3.4) and (3.5) completely describe the motion of the car. In practice there will be limits on the sizes of the forces u_1 and u_2 for obvious practical reasons. At this stage it's worth rewriting (3.4) and (3.5) in the form

$$\begin{bmatrix} dx_1/dt \\ dx_2/dt \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 1/m & -1/m \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \tag{3.6}$$

To construct (3.6) from its constituent parts, we have used the rule given in equations (1.59) and (1.60) in Chapter 1 for multiplying together a matrix and a vector. A compact matrix notation for (3.6) is

$$\frac{dx}{dt} = Ax + Bu \tag{3.7}$$

where in this example

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 1/m & -1/m \end{bmatrix}, \quad u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

Notice that the derivative dx/dt of the vector x consists of the vector with components dx_1/dt, dx_2/dt. In (3.7) we call x the *state vector*, and u the *control vector*, and their components the *state variables* and *control variables*, respectively. The state variables are so called because they tell us what the state of the system is.

Control problems could be as follows:

(i) Starting from rest at some initial point, that is $x_1(0) = 0$, $x_2(0) = 0$, find suitable functions $u_1(t)$, $u_2(t)$ so as to reach some given point $x_1(T) = a$ with zero final velocity $x_2(T) = 0$ in the least possible time T – perhaps a race from one set of green traffic lights to the next set at red!

(ii) Alternatively, the objective might be to reach $x_1(T) = a$ whilst consuming the least possible amount of fuel. To achieve this we would need to know how fuel consumption depends upon the velocity and acceleration, but clearly the optimum strategy would be quite different from that in (i), where instinct tells us that full throttle will be needed at least some of the time – but this will bring penalties in fuel consumption.

We'll be considering so-called *optimal* control problems like those in (i) and (ii), where the objective is to do something in the 'best possible' way, in Chapter 4.

To make the mathematical description of the car's motion more realistic we could take into account other factors such as road friction, wind resistance, engine speed, and so on. We could imagine the car to be travelling in one lane of a motorway, where an objective might be to make sure we keep within a reasonable distance from the vehicle in front. Indeed, the day might not be too far away when a metal strip embedded in the roadway will be used to control vehicles on motorways.

EXERCISE 3.2 List as many factors as you can think of which actually affect the motion of a car travelling along a road. Be imaginative: don't ignore factors which have only a very slight effect – they are all present in reality!

■ *EXAMPLE 3.3*

Mechanical systems involving springs are favourite models which can be used to illustrate ideas of control. Let's see what principles are involved. Consider first a carriage of mass m which runs along smooth, straight, horizontal rails, and is connected by a spring to a fixed vertical support as shown in Figure 3.3.

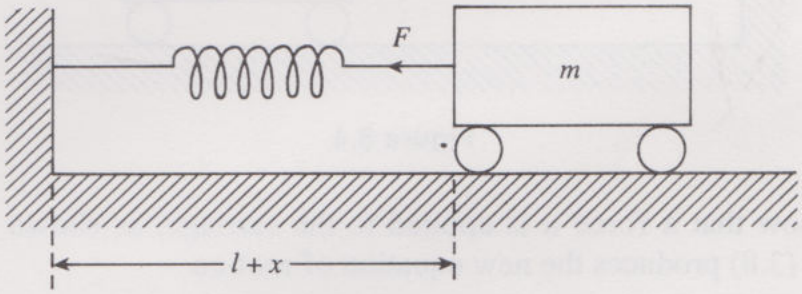


Figure 3.3

When the spring is neither extended nor compressed its length is l, it exerts no force and consequently there is no motion – the system is in equilibrium. Suppose the mass is pulled a distance x to the right, as shown in Figure 3.3. The spring is assumed to obey *Hooke's law*, which states that the force F it exerts on the mass is kx, where k is a constant for a given spring. Newton's law (3.3) tells us that the equation describing the motion of the mass when it is let go is

$$m\frac{d^2x}{dt^2} = -kx \tag{3.8}$$

since the acceleration is $f = d^2x/dt^2$. Notice that the force F exerted by the spring on the mass is in the opposite direction to that in which x is increasing, which accounts for the negative sign in (3.8). In other words, the spring pushes

back if compressed, and pulls back if stretched. If we write $k/m = w^2$ then (3.8) becomes

$$\frac{d^2x}{dt^2} = -w^2x \tag{3.9}$$

and you can easily verify by differentiating

$$x(t) = \alpha \cos wt + \beta \sin wt \tag{3.10}$$

twice that (3.10) is the solution of (3.9), where α and β are constants. The motion of the mass m described by (3.9) is oscillatory – that is, it moves backwards and forwards, with *period* $2\pi/w$. This means that the displacement of the mass from its equilibrium position at time t is the same at times $t + 2n\pi/w$, for $n = 1, 2, 3, \ldots$. This behaviour is called *simple harmonic motion.*

EXERCISE 3.3 Show that for $x(t)$ in (3.10)

$$x\left(t + \frac{2\pi n}{w}\right) = x(t)$$

for all positive integers n.

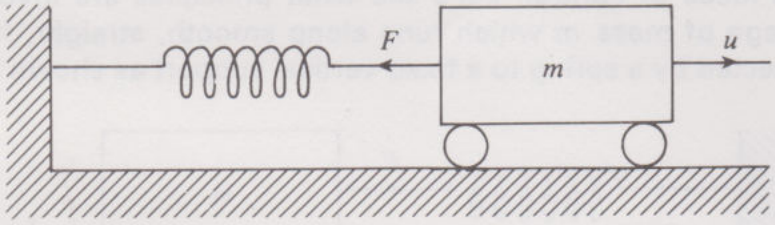


Figure 3.4

Suppose now that a force u is applied to the carriage, as shown in Figure 3.4. Adding this to (3.8) produces the new equation of motion

$$\frac{d^2x}{dt^2} = -w^2x + \frac{u}{m} \tag{3.11}$$

As in Example 3.2, let's now denote the distance x from equilibrium by x_1, and let the velocity dx/dt of the carriage at time t be x_2 as in (3.5). The single *second-order equation* (3.11) (so called because it contains a second-order derivative) can be converted into two first-order equations, just as we did in the previous example. We can write

$$\frac{d^2x}{dt^2} = \frac{d}{dt}\left(\frac{dx}{dt}\right) = \frac{dx_2}{dt}$$

so that (3.11) becomes

$$\frac{dx_2}{dt} = -w^2x_1 + \frac{u}{m}$$

Combining this equation with (3.5) gives us

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 1 \\ -w^2 & 0 \end{bmatrix}}_{A}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \underbrace{\begin{bmatrix} 0 \\ 1/m \end{bmatrix}}_{B} u \tag{3.12}$$

Again, this is in the form (3.7), where now there is just a single control variable u.

Let's now suppose a second carriage having mass m_2 is connected by another spring to the first carriage (now labelled m_1) as shown in Figure 3.5. An external force u is applied to the right-hand end.

The two springs have constants k_1 and k_2 respectively, and the forces they exert are denoted by F_1 and F_2. Let the displacements of the two carriages from equilibrium be x_1 and x_2. Suppose the first spring is extended by an amount x_1, so the force it exerts is $F_1 = k_1 x_1$. The net extension of the second spring is $x_2 - x_1$, since its right-hand end moves a distance x_2 and its left-hand end a distance x_1 (if $x_2 < x_1$ this means the second spring has a net compression, so the forces F_2 act in opposite directions to those shown in Figure 3.5). Newton's law of motion (3.3) now produces

$$\begin{aligned} m_1 \frac{\mathrm{d}^2 x_1}{\mathrm{d}t^2} &= F_2 - F_1 \\ &= k_2(x_2 - x_1) - k_1 x_1 \end{aligned} \tag{3.13}$$

for the first mass, and

$$\begin{aligned} m_2 \frac{\mathrm{d}^2 x_2}{\mathrm{d}t^2} &= -F_2 + u \\ &= -k_2(x_2 - x_1) + u \end{aligned} \tag{3.14}$$

for the second mass. Notice that the accelerations $\mathrm{d}^2x_1/\mathrm{d}t^2$ and $\mathrm{d}^2x_2/\mathrm{d}t^2$ carry the same signs as x_1 and x_2 respectively; forces which act in the opposing direction carry negative signs. Since there are now *two* second-order equations (3.13) and (3.14) we need to introduce *two* additional variables, which are the velocities

$$x_3 = \frac{\mathrm{d}x_1}{\mathrm{d}t}, \quad x_4 = \frac{\mathrm{d}x_2}{\mathrm{d}t} \tag{3.15}$$

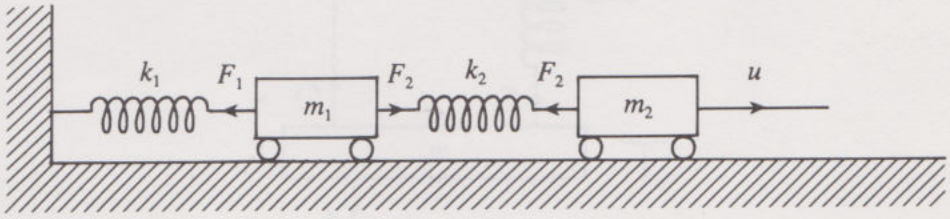


Figure 3.5

of the two carriages. Substituting into (3.13) gives us

$$m_1 \frac{dx_3}{dt} = -(k_1 + k_2)x_1 + k_2 x_2 \tag{3.16}$$

and from (3.14) we obtain

$$m_2 \frac{dx_4}{dt} = k_2 x_1 - k_2 x_2 + u \tag{3.17}$$

You should verify that the equations (3.15), (3.16) and (3.17) can be written in the combined form

$$\frac{d}{dt}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -(k_1+k_2)/m_1 & k_2/m_1 & 0 & 0 \\ k_2/m_2 & -k_2/m_1 & 0 & 0 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}}_{x} + \underbrace{\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1/m_2 \end{bmatrix}}_{B} u \tag{3.18}$$

which again gives us (3.7) with an appropriate interpretation of x, A, B and u. Indeed, this expression (3.7) is our general model of a linear control system when time is measured continuously. In general if there are n state variables $x_1, \ldots, x_n$ and m control variables $u_1, u_2, \ldots, u_m$ then A is a square $n \times n$ matrix and B is an $n \times m$ matrix.

EXERCISE 3.4 Consider the mechanical system shown in Figure 3.6. This consists of two masses m_1 and m_2 which hang vertically from a fixed support to which they are

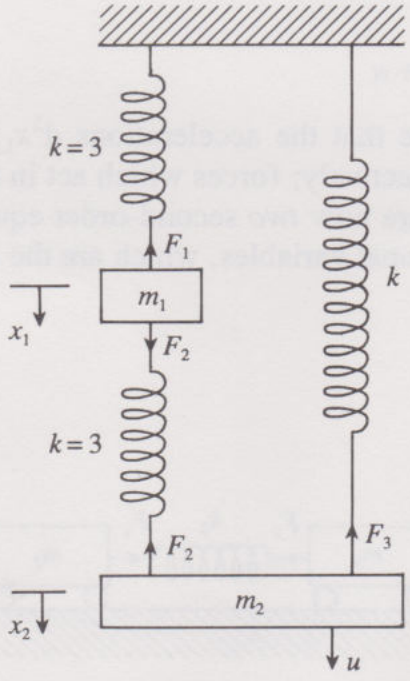


Figure 3.6

connected by three springs as shown. The forces exerted by the springs are F_1, F_2 and F_3 when the downwards displacements of the masses from the equilibrium position are x_1 and x_2. The spring constants have the values shown in the figure. If $m_1 = 1$, $m_2 = 4$, write down Newton's equation of motion for each of the two masses (you do not have to take gravitational forces into account). If the velocities are $dx_1/dt = x_3$, $dx_2/dt = x_4$ show that the equations can be written in the form (3.7) as follows:

$$\frac{d}{dt}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -6 & 3 & 0 & 0 \\ \frac{3}{4} & -(3+k)/4 & 0 & 0 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ \frac{1}{4} \end{bmatrix} u \tag{3.19}$$

■ *EXAMPLE 3.4*

In the mechanical system shown in Figure 3.3 we ignored all frictional forces so that when the mass is set moving it performs oscillatory motions which continue indefinitely. In practice there will be resisting or *damping* forces which oppose the motion. For example, if a car was supported only on springs then after hitting a bump in the road it would bounce up and down in a highly uncomfortable way. For this reason the suspension system contains damping elements called 'shock absorbers' which cause the oscillatory motion induced by going over a bump to die away rapidly. Of course, in practice the elasticity of the tyres must also be taken in account.

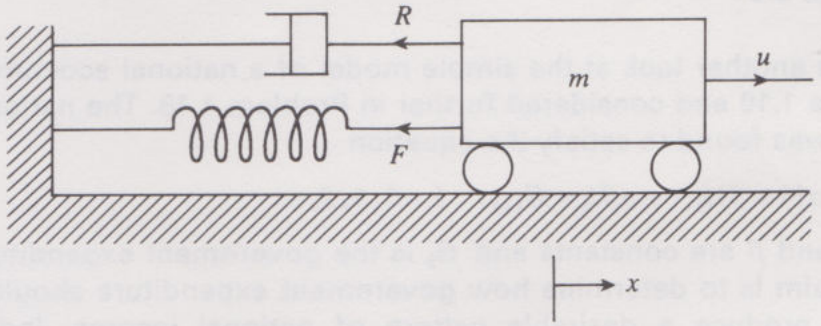


Figure 3.7

In Figure 3.7 the damper is shown as exerting a force R opposing the motion (the diagrammatic representation is of a piston in a cylinder). It is a reasonable approximation to reality to assume that the damping force is proportional to the relative *velocity*, so that here $R = p\,dx/dt$, where p (> 0) is the *damping constant*. Ignore for the present the force u shown in Figure 3.7. The equation of motion (3.8) now becomes

$$m\frac{d^2x}{dt^2} = -kx - p\frac{dx}{dt} \tag{3.20}$$

EXERCISE 3.5 Consider equation (3.20) with $m=1$, $p=4$ and $k=5$. Verify that

$$x(t)=e^{-2t}(\alpha \cos t+\beta \sin t) \tag{3.21}$$

where α and β are arbitrary constants, is the general solution of the equation by finding dx/dt and d^2x/dt^2 and substituting into (3.20).

In the preceding exercise it follows from (3.21) that $x(t) \to 0$ as $t \to \infty$, since the exponential term certainly has this property, and the term $\alpha \cos t+\beta \sin t$ can never exceed $|\alpha|+|\beta|$ because $|\cos t| \leqslant 1$, $|\sin t| \leqslant 1$ ($|\alpha|$ is the modulus of α, defined in Section 1.2, Chapter 1). In fact it can be shown in general that the solution of (3.20) always tends to zero as t becomes larger and larger provided k and p are both positive (see Problem 3.1). In other words, for a damped system like that in Figure 3.7, oscillations always die away and the system returns to rest after a sufficiently long time; the larger the value of p, the more rapidly do oscillations decay.

EXERCISE 3.6 Suppose that a force $u(t)$ is applied to the carriage as shown in Figure 3.7, so that a term u is added to the right-hand side in equation (3.20). Using the numerical values of m, p and k in Exercise 3.5, by taking $x=x_1$ and $dx/dt=x_2$ show that (3.20) can be written in the form (3.7) with

$$x=\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad A=\begin{bmatrix} 0 & 1 \\ -5 & -4 \end{bmatrix}, \quad B=\begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{3.22}$$

■ *EXAMPLE 3.5*

Let's have another look at the simple model of a national economy introduced in Exercise 1.10 and considered further in Problem 1.18. The national income I_k at year k was found to satisfy the equation

$$I_{k+2}-\alpha(1+\beta)I_{k+1}+\alpha\beta I_k=G_{k+2}, \quad k=0, 1, 2, \ldots \tag{3.23}$$

where α and β are constants and G_k is the government expenditure in year k. Here the aim is to determine how government expenditure should be planned so as to produce a desirable pattern of national income. Indeed, finance ministers around the world would dearly like to know how this could be achieved – the 'science' of mathematical economics is still in its infancy. For example, it was found in Problem 1.18 that (under the assumptions of the model as set out in Exercise 1.10) with government expenditure kept constant the national income behaves in an oscillatory fashion, but eventually settles down to twice government expenditure – a result which would not be predicted by 'common sense' arguments.

The equation (3.23) has been reintroduced to show you that models of control systems can be either in the form of differential equations like those in Examples 3.2, 3.3 and 3.4 or in the form of difference equations like (3.1) or (3.23). The

general matrix description in the difference equation case, as compared with (3.7) for the differential equation case, is

$$x(k+1) = Ax(k) + Bu(k), \quad k = 0, 1, 2, \ldots \tag{3.24}$$

This is just the matrix model (1.58) with the control term $Bu(k)$ added on.

EXERCISE 3.7 By setting $x_1(k) = I_k$, $x_2(k) = I_{k+1}$ show that (3.23) can be expressed in the form (3.24) with

$$x(k) = \begin{bmatrix} x_1(k) \\ x_2(k) \end{bmatrix}, \quad A = \begin{bmatrix} 0 & 1 \\ -\alpha\beta & \alpha(1+\beta) \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad u(k) = G_{k+2}$$

■ *EXAMPLE 3.6*

If you take a dose of medicine it first enters your gastrointestinal tract. From there it is distributed throughout your bloodstream to be metabolized and eventually eliminated. At a particular instant of time t, let x_1 be the mass of drug in the gastrointestinal tract, let x_2 be the mass of drug in the bloodstream, and let u be the rate at which the drug is taken (these variables are all functions of t). Then the process is described by

$$\frac{dx_1}{dt} = -k_1 x_1 + u$$

rate of increase of drug in tract — rate at which drug passes to bloodstream — rate of drug ingestion

and

$$\frac{dx_2}{dt} = k_1 x_1 - k_2 x_2$$

rate of increase of drug in bloodstream — rate of receipt of drug into bloodstream — rate of drug excretion

where k_1 and k_2 are positive constants depending upon characteristics of the body. If these equations are written in the form (3.7) then you should verify that

$$A = \begin{bmatrix} -k_1 & 0 \\ k_1 & -k_2 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

The objective is to determine the rate u at which the drug should be given to a patient so as to control the amount of the drug in the bloodstream according to medically desirable levels.

EXERCISE 3.8 Suppose in the situation described in Example 3.6 there are initial amounts of $x_1(0) = a$ and $x_2(0) = b$ of drug present. If no further doses of drug are administered, verify that the equations in Example 3.6 are satisfied by

$$x_1(t) = ae^{-k_1 t}$$

$$x_2(t) = be^{-k_2 t} + \frac{k_1 a}{k_1 - k_2}(e^{-k_2 t} - e^{-k_1 t})$$

assuming $k_1 \neq k_2$. This shows that $x_1(t) \to 0$, $x_2(t) \to 0$ as $t \to \infty$.

■ *EXAMPLE 3.7*

An oven has a rectangular cross-section as shown in Figure 3.8.

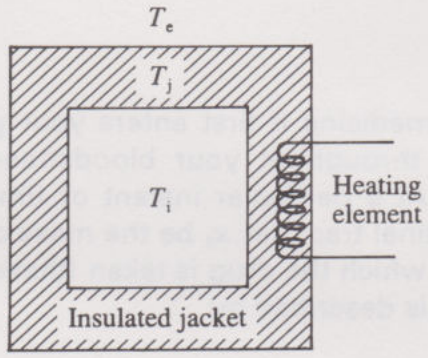


Figure 3.8

The objective is to control the temperature T_i of the interior of the oven by varying the heat input u to the jacket, which is insulated. The energy is supplied by means of an electric heating element. Let T_j and T_e denote the temperatures of the jacket and external surroundings respectively. The oven interior gains heat solely by radiation from the inside surface of the jacket. The rate at which this energy is radiated is proportional to the temperature difference $T_j - T_i$, and also depends upon the interior surface area a_i of the jacket. The rate of increase of heat energy of the oven interior is therefore given by

$$c_i \frac{dT_i}{dt} = a_i r_i (T_j - T_i) \tag{3.25}$$

where c_i is the heat capacity of the oven interior, and r_i is called the *radiation coefficient* of the surface.

In a similar way the jacket also radiates heat to the external surroundings, so the 'heat balance' equation for the jacket is

$$c_j \frac{dT_j}{dt} = -a_i r_i (T_j - T_i) - a_e r_e (T_j - T_e) + u \tag{3.26}$$

rate of increase of heat energy of jacket | rate of heat loss to oven interior | rate of heat loss to external surroundings | heat input

where c_j is the heat capacity of the jacket, a_e is the area of the external surface of the jacket and r_e is its radiation coefficient. It is assumed that T_e is constant. The heat input u to the jacket is varied by altering the current through the coil. As this input is not applied directly to the oven interior, a first question to ask is whether it is indeed possible to maintain the oven interior temperature T_i at any desired level by appropriately altering u. This is an illustration of what is meant by 'controllability' of a system, and will be investigated in Section 3.2 (in particular, see Example 3.10). A second question is whether the value of the oven interior temperature T_i can be determined even if it cannot be measured directly – perhaps we can only measure the jacket temperature T_j. This is an example of the problem of 'observability' of a system, and will be studied in Section 3.3 (in particular, see Example 3.13). We shall also find that although the two questions seem to involve quite different aspects of physical reality, they are actually closely related in mathematical terms.

EXERCISE 3.9 When a deep-sea diver is brought up to the surface this is done by attaching the diver to a cable which is operated by a winch. Assuming that the motion takes place entirely vertically, then at a depth h below the surface the equation of motion is obtained from Newton's law as

$$m\frac{d^2h}{dt^2} = (mg - \rho v) - \mu\frac{dh}{dt} - f$$

where m is the mass of the diver, v is the volume of the diver, ρ is the density of water, μ is a positive drag coefficient, f is the force exerted by the cable and g is the gravitational constant. Let p denote the internal body pressure of the diver relative to atmospheric pressure at sea level. It is important for the health of the diver to avoid large changes in the body pressure p whilst being raised to the surface by the cable. It can be shown that

$$\frac{dp}{dt} = k(\rho h - p)$$

where k is a positive constant characterizing the body tissue. Define as state variables

$$x_1 = h, \quad x_2 = \frac{dh}{dt}, \quad x_3 = p$$

and let the control variable be

$$u = \frac{mg - \rho v - f}{m}$$

Write the equations in the matrix form (3.7) where x is the vector with components x_1, x_2, x_3.

EXERCISE 3.10 An overhead crane of mass M moves along a horizontal track, and its distance at time t from a fixed reference point is s. A grab of mass m is attached to the crane by a rod whose mass can be neglected, as shown in Figure 3.9. The angle θ

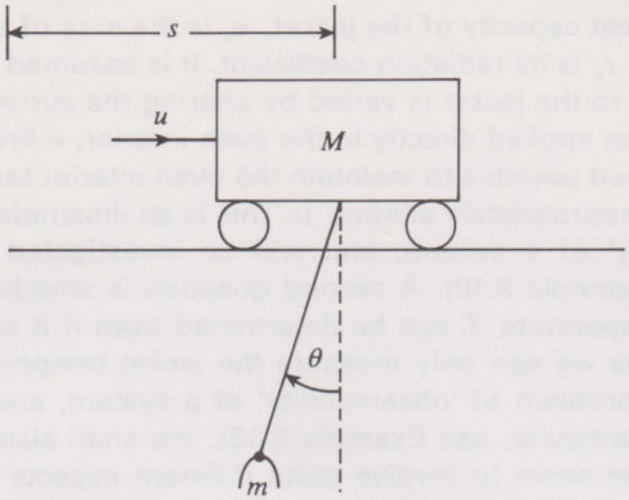


Figure 3.9

made by this rod to the vertical is assumed small, in which case the equations of motion turn out to be

$$M\frac{d^2\theta}{dt^2} + (m+M)g\theta + u = 0$$

$$M\frac{d^2s}{dt^2} - mg\theta - u = 0$$

where u is a control force. Taking $M=1$, $m=0.1$ write these equations in the matrix form (3.7) by taking as state variables

$$x_1 = \theta, \quad x_2 = \frac{d\theta}{dt}, \quad x_3 = s, \quad x_4 = \frac{ds}{dt}$$

EXERCISE 3.11 Return to the model of the buffalo population in the American west described in Problem 1.28. It was found that

$$F_{k+2} = 0.95F_{k+1} + 0.12F_k \tag{3.27}$$

$$M_{k+2} = 0.95M_{k+1} + 0.14F_k, \quad k = 0, 1, 2, \ldots \tag{3.28}$$

where F_k and M_k are the numbers of female and male buffalo at the start of year k, where $k=0$ corresponds to 1830. You were asked to show that under natural conditions the numbers of animals would continue to grow indefinitely. In fact owing to indiscriminate slaughter by white settlers who were only interested in buffalo hides, the number of animals was reduced from an estimated 60 million in 1830 to just a few hundred only 60 years later. Suppose that a policy of strictly controlled slaughter had been adopted, whereby a number of adult females were killed for food each year. This is equivalent to an extra control term $-u(k)$ on the right-hand side of equation (3.27). Define the state variables

$$x_1(k) = F_k, \quad x_2(k) = F_{k+1}, \quad x_3(k) = M_k, \quad x_4(k) = M_{k+1}$$

and hence write (3.27) and (3.28) in the form (3.24).

3.2 CONTROLLABILITY

Politicians the world over would desperately like to be able to control the economy of their country in ways they think desirable – especially ways which appeal to voters! For example, to keep down the rate of inflation, to reduce unemployment and to raise living standards are all admirable but somehow elusive targets. Fashionable ideas change: one party claims that controlling the supply of money is the answer, another relies on manipulating the interest rate, whilst a third calls for more government investment. There is a tendency for politicians to use jargon which confuses the electorate, such as 'the unemployment rate is a lagging indicator', but this is really a smokescreen designed to obscure the real lack of understanding of how to control an economy. In fact the first question to be asked is: can the results we wish to obtain be actually achieved by altering the factors we have selected? For example, can inflation be controlled by altering interest rates? Can unemployment be brought down by reducing taxes? Can living standards be raised by increasing investment? All these are questions of *controllability*; can a system be compelled to behave in a certain way by altering the variables which we have selected as the inputs? If the answer is 'yes' then we can go about devising suitable control schemes which will do the job; but if the answer is 'no' then any hope of constructing a successful strategy for control is doomed to failure. In this latter case we must alter the way in which control is applied, perhaps by selecting a different set of control variables, so as to make sure we get a set-up which is *controllable*.

To investigate controllability it's easier to begin with difference equations.

■ *EXAMPLE 3.8*

Let's look at a discrete time model in the form (3.24), namely

$$x(k+1) = Ax(k) + Bu(k), \quad k = 0, 1, 2, \ldots \tag{3.29}$$

and suppose for simplicity there are *two* state variables $x_1(k)$ and $x_2(k)$, and a *single* control variable. This means that in (3.29) A is a 2×2 matrix and B is a 2×1 column vector which we will denote by b to emphasize that it is a *vector*. Suppose our objective is to drive the system from a given initial state $x(0)$ to a given final state x_f in two units of time, that is we want to make $x(2) = x_f$. Setting $k = 0$ in (3.29) gives us

$$x(1) = Ax(0) + bu(0)$$

and setting $k = 1$ in (3.29) gives

$$\begin{aligned} x(2) &= Ax(1) + bu(1) \\ &= A[Ax(0) + bu(0)] + bu(1) \\ &= A^2x(0) + Abu(0) + bu(1) \end{aligned} \tag{3.30}$$

Our control problem is therefore to determine the values of the control $u(0)$ and $u(1)$ such that the expression in (3.30) is equal to any given final state x_f.

Let's write this out as

$$x_f = A^2x(0) + [b, Ab]\begin{bmatrix} u(1) \\ u(0) \end{bmatrix} \tag{3.31}$$

You should notice carefully how we were able to convert the terms $Abu(0) + bu(1)$ in (3.30) into the expression on the right in (3.31). It's worth writing it out in full detail just this once: supposing that

$$b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \quad Ab = \begin{bmatrix} b_3 \\ b_4 \end{bmatrix}$$

then the term in (3.31) is

$$\begin{aligned} [b, Ab]\begin{bmatrix} u(1) \\ u(0) \end{bmatrix} &= \begin{bmatrix} b_1 & b_3 \\ b_2 & b_4 \end{bmatrix}\begin{bmatrix} u(1) \\ u(0) \end{bmatrix} \\ &= \begin{bmatrix} b_1u(1) + b_3u(0) \\ b_2u(1) + b_4u(0) \end{bmatrix} \\ &= \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}u(1) + \begin{bmatrix} b_3 \\ b_4 \end{bmatrix}u(0) \\ &= bu(1) + Abu(0) \end{aligned}$$

which agrees with the original expression in (3.30).

We can arrange (3.31) into the form

$$[b, Ab]\begin{bmatrix} u(1) \\ u(0) \end{bmatrix} = x_f - A^2x(0)$$

and then *invert* the matrix $U = [b, Ab]$ to obtain

$$\begin{bmatrix} u(1) \\ u(0) \end{bmatrix} = U^{-1}[x_f - A^2x(0)] \tag{3.32}$$

provided that the matrix U^{-1} exists. If this is the case then *whatever* the final state x_f we can use (3.32) to compute control values $u(0)$ and $u(1)$ which will indeed steer the system from *any* initial state $x(0)$ to $x(2) = x_f$.

We have already encountered the concept of the *inverse matrix* U^{-1} in Section 1.4, Chapter 1. Recall that

$$UU^{-1} = U^{-1}U = I$$

where I is the unit matrix having ones along the principal diagonal. We gave the formula for the inverse of a 2×2 matrix in (1.80), repeated here for convenience:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \tag{3.33}$$

provided the *determinant* $ad - bc \neq 0$. This latter condition is required for a matrix to have an inverse, in which case the matrix is called *non-singular* (otherwise it is

singular). We can therefore say that the system (3.29) is *controllable* provided the 2×2 *controllability matrix* $U = [b, Ab]$ is non-singular, that is

$$\det U = ad - bc \neq 0 \tag{3.34}$$

EXERCISE 3.12 Consider a system described by the difference equation

$$x(k+1) = \begin{bmatrix} 0 & 1 \\ -6 & 5 \end{bmatrix} x(k), \quad k = 0, 1, 2, \ldots \tag{3.35}$$

(a) If

$$x(0) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

show that

$$x(2) = \begin{bmatrix} -1 \\ -11 \end{bmatrix}$$

(b) If a control term

$$\begin{bmatrix} 0 \\ 1 \end{bmatrix} u(k)$$

is added onto the right-hand side in (3.35), determine the controllability matrix U, and show that the system is controllable.

(c) Use (3.32) to find the values of $u(0)$ and $u(1)$ which send the system from the initial state $x(0)$ in part (a) to the final state

$$x(2) = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

EXERCISE 3.13 Consider the system described by

$$x(k+1) = \begin{bmatrix} 1 & 2 \\ -1 & 4 \end{bmatrix} x(k) + \begin{bmatrix} \alpha \\ 1 \end{bmatrix} u(k), \quad k = 0, 1, 2, \ldots$$

Find for what values of the parameter α the system is *not* controllable.

Let's return to (3.29) and suppose that there are now *three* state variables and a single control variable, so we have

$$x(k+1) = Ax(k) + bu(k), \quad k = 0, 1, 2, 3, \ldots \tag{3.36}$$

where A is a 3×3 matrix and b is a 3×1 column vector. As before, set $k = 0$, 1 in (3.36) to get the expression for $x(2)$ in (3.30), but we now go one step further and take $k = 2$ which gives from (3.36)

$$\begin{aligned} x(3) &= Ax(2) + bu(2) \\ &= A^3x(0) + A^2bu(0) + Abu(1) + bu(2), \quad \text{using (3.30)} \\ &= A^3x(0) + \underbrace{[b, Ab, A^2b]}_{U} \begin{bmatrix} u(2) \\ u(1) \\ u(0) \end{bmatrix} \end{aligned} \tag{3.37}$$

Using the same argument as for the 2×2 case, it follows that we can obtain a control sequence $u(0)$, $u(1)$, $u(2)$ which sends the system to *any* final state $x(3) = x_f$ provided U^{-1} exists. Hence for the system (3.36) to be controllable we must have $\det U \neq 0$ exactly as in (3.34), but now U is the 3×3 *controllability matrix*

$$U = [b, Ab, A^2b] \tag{3.38}$$

Notice that to compute the third column A^2b in (3.38) we do *not* need A^2. First work out Ab, and then use

$$A^2b = A(Ab)$$

A formula for evaluating a 3×3 determinant in terms of three 2×2 determinants was given in Chapter 1, equations (1.86) and (1.87).

■ *EXAMPLE 3.9*

Let's investigate whether the system described by

$$x(k+1) = \underbrace{\begin{bmatrix} 1 & -1 & 2 \\ 3 & 0 & 4 \\ -5 & 6 & -2 \end{bmatrix}}_{A} x(k) + \underbrace{\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}}_{b} u(k), \quad k = 0, 1, 2, \ldots \tag{3.39}$$

is controllable. Applying the usual rule (set out in (1.60)) for multiplying together a matrix and a vector we get

$$Ab = \begin{bmatrix} 1 \times 1 - 1 \times 0 + 2 \times 1 \\ 3 \times 1 + 0 \times 0 + 4 \times 1 \\ -5 \times 1 + 6 \times 0 - 2 \times 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 7 \\ -7 \end{bmatrix}$$

and similarly

$$\begin{aligned} A^2b &= A(Ab) \\ &= A\begin{bmatrix} 3 \\ 7 \\ -7 \end{bmatrix} = \begin{bmatrix} 1 \times 3 - 1 \times 7 - 2 \times 7 \\ 3 \times 3 + 0 \times 7 - 4 \times 7 \\ -5 \times 3 + 6 \times 7 + 2 \times 7 \end{bmatrix} = \begin{bmatrix} -18 \\ -19 \\ 41 \end{bmatrix} \end{aligned}$$

The controllability matrix in (3.38) is therefore obtained by writing the three columns, b, Ab, A^2b side by side to give

$$U = \begin{bmatrix} 1 & 3 & -18 \\ 0 & 7 & -19 \\ 1 & -7 & 41 \end{bmatrix}$$

and we need to evaluate the determinant of U using (1.86). This gives

$$\begin{aligned} \det U &= 1\begin{vmatrix} 7 & -19 \\ -7 & 41 \end{vmatrix} - 3\begin{vmatrix} 0 & -19 \\ 1 & 41 \end{vmatrix} - 18\begin{vmatrix} 0 & 7 \\ 1 & -7 \end{vmatrix} \\ &= (7 \times 41 - 7 \times 19) - 3(1 \times 19) - 18(-1 \times 7) \\ &= 154 - 57 + 126 = 223 \end{aligned}$$

which is non-zero, showing that (3.39) is controllable.

EXERCISE 3.14 Test the system (3.36) for controllability when

$$A=\begin{bmatrix}1 & 3 & -2\\ 1 & 2 & -1\\ 1 & 1 & 0\end{bmatrix},\quad b=\begin{bmatrix}1\\0\\0\end{bmatrix}$$

EXERCISE 3.15 Use (3.37) to show that for a controllable system (3.36) the control sequence $u(0)$, $u(1)$, $u(2)$ which sends the system from $x(0)$ to any given state x_f after three units of time is

$$\begin{bmatrix}u(2)\\u(1)\\u(0)\end{bmatrix}=U^{-1}[x_f-A^3x(0)] \tag{3.40}$$

where U^{-1} is the inverse of the controllability matrix (3.38).

The preceding exercise shows that for a controllable system with *three* state variables we can get to any final state in *three* units of time; previously we saw in (3.32) that with *two* state variables, *two* units of time are required. It therefore comes as no surprise that when there are n state variables and a single control variable, so in (3.36) A is $n\times n$ and b is $n\times 1$, the following result holds:

Provided the $n\times n$ *controllability matrix* defined by

$$U=[b, Ab, A^2b, A^3b, \ldots, A^{n-1}b] \tag{3.41}$$

is non-singular (i.e. $\det U\neq 0$), then a control sequence $u(0)$, $u(1)$, $u(2)$, ..., $u(n-1)$ can be found which drives the system (3.36) from any initial state to any final state in n units of time.

Notice that, as was pointed out for $n=3$, it is *not* necessary to work out powers of A in order to compute the columns of U in (3.41). Each column is obtained by multiplying the preceding one by A as follows:

$$A^2b=A(Ab),\quad A^3b=A(A^2b),\quad A^4b=A(A^3b),\quad \ldots \tag{3.42}$$

Of course we haven't yet covered the evaluation of determinants when $n>3$. We'll defer this until Section 3.5, where we'll also see what happens when there are several control variables. In fact, even for the case of 3×3 determinants where we have so far used the formula (1.86) we'll see that an improved method of evaluation is available.

Let's now turn to the case when time is regarded as continuous, so our linear system model consists of the matrix *differential* equation in (3.7), namely

$$\frac{dx}{dt}=Ax+bu \tag{3.43}$$

where again we are assuming at present that there is a single control variable, so b is an $n\times 1$ column vector and A is an $n\times n$ matrix, and $x(t)$, $u(t)$ are continuous functions of time t. It very often turns out that it is more complicated to deal with models using differential equations than those using difference equations, and this is

certainly true when we try to establish the controllability condition for (3.43). We therefore shan't attempt to prove this, but simply state the result – which actually turns out to be the *same* in mathematical terms!

The system (3.43) is *controllable*, in the sense that there exists a control function $u(t)$ which transfers the system from any initial state $x(0)$ to any final state x_f in a finite time, provided the controllability matrix U defined in (3.41) is non-singular. However, this time there are no simple expressions like (3.32) or (3.40) for controls which will achieve the desired result.

■ *EXAMPLE 3.10*

Return to the electrically heated oven described in Example 3.7. Define as the state variables the *excesses* of temperature over that of the surroundings T_e, that is

$$x_1 = T_i - T_e, \quad x_2 = T_j - T_e$$

Since T_e is assumed constant we have

$$\frac{dx_1}{dt} = \frac{dT_i}{dt}, \quad \frac{dx_2}{dt} = \frac{dT_j}{dt}$$

so (3.25) and (3.26) become

$$c_i \frac{dx_1}{dt} = a_i r_i (x_2 - x_1)$$

$$c_j \frac{dx_2}{dt} = -a_i r_i (x_2 - x_1) - a_e r_e x_2 + u$$

where we have used the fact that $x_2 - x_1 = T_j - T_i$. To avoid messy algebra, suppose $c_i = 1$, $c_j = 2$, $a_i = 10$, $a_e = 30$, $r_i = 2$, $r_e = 1$.

You should check that we can then write these equations in the matrix form

$$\frac{d}{dt}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \underbrace{\begin{bmatrix} -20 & 20 \\ 10 & -25 \end{bmatrix}}_{A}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \underbrace{\begin{bmatrix} 0 \\ \frac{1}{2} \end{bmatrix}}_{b} u$$

The controllability matrix is

$$U = [b, Ab]$$

$$= \begin{bmatrix} 0 & 10 \\ \frac{1}{2} & -\frac{25}{2} \end{bmatrix}$$

and

$$\det U = -10 \times \tfrac{1}{2} = -5 \neq 0$$

We therefore conclude that the system is controllable; that is, we can indeed raise the temperature of the oven interior from any initial value to any final

value in a finite time merely by altering the current through the heating element in the jacket.

EXERCISE 3.16 Verify that (a) the mechanical system in (3.12) is controllable, and (b) the drug model in Example 3.6 is controllable.

EXERCISE 3.17 Determine for what values of the real parameter a the system

$$\frac{dx}{dt} = \begin{bmatrix} 2 & a-3 \\ 0 & 2 \end{bmatrix} x + \begin{bmatrix} 1 \\ a^2 - a \end{bmatrix} u$$

is *not* controllable.

EXERCISE 3.18 In Chapter 1 we described in Example 1.4 Fibonacci's model for a population of rabbits. We saw that if rabbits live in 'paradise' – an infinitely large green island with no other animals or diseases – then their numbers will increase without limit. This time let's use a continuous time model: denote by $x_1(t)$ the number of rabbits at time t. Suppose that growth is *exponential*, that is

$$x_1(t) = e^{at} x_1(0) \tag{3.44}$$

for some *positive* constant a, where $x_1(0)$ is the initial number of rabbits when counting begins at $t = 0$. Differentiating (3.44) with respect to t gives

$$\frac{dx_1}{dt} = ae^{at} x_1(0)$$
$$= ax_1 \tag{3.45}$$

Unfortunately for the rabbits, foxes are introduced onto the island. These carnivorous animals feed on the rabbits; indeed, if there were *no* rabbits on the island the foxes would simply die out, again at an exponential rate, so that exactly as in (3.45) we would have

$$x_2(t) = e^{-bt} x_2(0), \quad \frac{dx_2}{dt} = -bx_2 \tag{3.46}$$

where $x_2(t)$ is the number of foxes at time t and b is another positive constant. We now assume that when rabbits and foxes cohabit on the island:

(i) the rate of growth of the rabbit population is *reduced* by an amount proportional to the number of foxes (i.e. by an amount $-cx_2$);
(ii) the rate of growth of the fox population is *increased* by an amount proportional to the number of rabbits (i.e. by an amount dx_1).

These assumptions mean that (3.45) and (3.46) are replaced respectively by

$$\frac{dx_1}{dt} = ax_1 - cx_2 \tag{3.47}$$

$$\frac{dx_2}{dt} = dx_1 - bx_2 \tag{3.48}$$

where c and d are also positive constants. In matrix form (3.47) and (3.48) can be written as

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \underbrace{\begin{bmatrix} a & -c \\ d & -b \end{bmatrix}}_{A}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{3.49}$$

Depending upon the values of the parameters a, b, c and d, the numbers of animals can increase without limit as $t \to \infty$. It isdecided to attempt to control the environment by introducing a disease which is fatal to rabbits but does not harm foxes. This has the effect of adding a control term u onto the right-hand side in (3.47). Test whether the resulting system is controllable – that is, whether applying control in this way will allow the numbers of both rabbits *and* foxes to be brought to desired levels.

3.3 OBSERVABILITY

The state variables tell us everything there is to know about a system at any particular time. For example, if a car is being driven along a road the state variables could include the speed, acceleration and position of the vehicle, its mass (which is decreasing due to consumption of fuel), the engine temperature and oil pressure, and so on. The values of all these things define the *state* of the car, and knowing what's happening enables us to drive accordingly. In fact, it's still possible to drive the car perfectly well without knowing, for example, what the oil pressure is – and, indeed, few cars have an oil pressure gauge these days. So for reasons of cost or practicability it's very often the case that only certain variables, called *outputs*, are monitored and their values used to control the system. The problem of *observability* is whether it's possible to determine the state of a system from a knowledge of its outputs (it is assumed that we know what inputs we are using). If the answer turns out to be 'no', so the system is *not observable*, then a different set of outputs will have to be selected for monitoring.

Consider as another example the economy of a country: finding out its state is exceedingly difficult. The actual number of state variables is immense – for example, the daily income and expenditure of every individual is a crucial element of the total information. Clearly it is completely impractical to monitor all these constituents of the overall economic picture. Instead, key indicators such as the levels of imports and exports, the volume of manufacturing output, the amount of money in circulation, and so on, are measured and those in charge of economic affairs then try to determine what the true state of the economy is.

To keep things simple we'll just look now at the case where there is a *single* output variable y, and leave the situation where there are several outputs until Section 3.5. We assume that the output y is a *linear combination* of the states. This means (as in equation (1.67), Chapter 1) that

$$y = c_1x_1 + c_2x_2 + \cdots + c_nx_n \tag{3.50}$$

where $c_1, c_2, \ldots, c_n$ are constants. We can also write (3.50) as

$$y = cx \tag{3.51}$$

where c is the $1 \times n$ row vector with components $c_1, \ldots, c_n$ and x is the usual $n \times 1$ state vector with components $x_1, \ldots, x_n$. The expression cx in (3.51) is called the *scalar product* of the two vectors c and x. Once again it turns out to be easier to begin with the discrete time case, so in (3.50) we understand that each of the variables $y(k), x_1(k), \ldots, x_n(k)$ is defined for $k = 0, 1, 2, \ldots$. Recall that our model is the set of difference equations (3.29). We know completely what inputs we are using (i.e. the control sequence $u(0), u(1), u(2), \ldots$) so whatever these are won't affect the observability question. It's obviously going to simplify matters if we suppose our input is zero, that is $u(k) = 0, k \geqslant 0$.

The equations describing the system are therefore just

$$x(k+1) = Ax(k), \quad k = 0, 1, 2, \ldots \tag{3.52}$$

Setting $k = 0$ in (3.51) gives

$$y(0) = cx(0) \tag{3.53}$$

and similarly $k = 1$ produces

$$y(1) = cx(1) = cAx(0) \tag{3.54}$$

since $x(1) = Ax(0)$ from (3.52). Continuing this process gives

$$y(2) = cx(2) = cAx(1) = cA^2x(0) \tag{3.55}$$

and

$$y(3) = cA^3x(0), \ldots, y(n-1) = cA^{n-1}x(0) \tag{3.56}$$

We can combine together the equations (3.53) to (3.56) to produce

$$\begin{bmatrix} y(0) \\ y(1) \\ y(2) \\ \vdots \\ y(n-1) \end{bmatrix} = \begin{bmatrix} c \\ cA \\ cA^2 \\ \vdots \\ cA^{n-1} \end{bmatrix} x(0) \tag{3.57}$$

$$= Vx(0) \tag{3.58}$$

where V is the $n \times n$ matrix on the right-hand side of (3.57) with rows c, cA, $cA^2, \ldots, cA^{n-1}$. Provided this matrix V has an inverse V^{-1} we can solve (3.58) to produce

$$x(0) = V^{-1} \begin{bmatrix} y(0) \\ y(1) \\ \vdots \\ y(n-1) \end{bmatrix} \tag{3.59}$$

Equation (3.59) expresses the initial state of the system in terms of the outputs at times $0, 1, 2, \ldots, n-1$. Once $x(0)$ is known we can find $x(1)$, $x(2), \ldots$ by using (3.52). In fact, we saw in Chapter 1, equation (1.61), that $x(k) = A^k x(0)$, $k \geqslant 1$. In other words, we can indeed determine the state of the system by measuring the outputs, so the system is *observable* provided the *observability matrix* V is non-singular, that is

$$\det V \neq 0 \tag{3.60}$$

■ EXAMPLE 3.11

Let's return to the discrete system of Example 3.9, where A is the 3×3 matrix in (3.39). Suppose that the output variable which is measured is

$$y(k) = 2x_1(k) - x_3(k) \tag{3.61}$$

so that $c = [2, 0, -1]$. The observability matrix V has first row equal to c, second row

$$cA = [2, 0, -1]\begin{bmatrix} 1 & -1 & 2 \\ 3 & 0 & 4 \\ -5 & 6 & -2 \end{bmatrix}$$

$$= [2 \times 1 + 0 \times 3 - 1 \times -5,\ 2 \times -1 + 0 \times 0 - 1 \times 6,\ 2 \times 2 + 0 \times 4 - 1 \times -2]$$

$$= [7, -8, 6]$$

and third row

$$cA^2 = (cA)A$$

$$= [7, -8, 6]A$$

$$= [-47, 29, -30]$$

Notice that as in the case of constructing the columns of a controllability matrix via (3.42), we do *not* compute powers of A, but use the expressions

$$cA^2 = (cA)A, \quad cA^3 = (cA^2)A, \ldots$$

The matrix V is therefore

$$V = \begin{bmatrix} 2 & 0 & -1 \\ 7 & -8 & 6 \\ -47 & 29 & -30 \end{bmatrix}$$

and using (1.86) the determinant of V is

$$\det V = 2\begin{vmatrix} -8 & 6 \\ 29 & -30 \end{vmatrix} - 0\begin{vmatrix} 7 & 6 \\ -47 & -30 \end{vmatrix} - 1\begin{vmatrix} 7 & -8 \\ -47 & 29 \end{vmatrix}$$

$$= 2[(-8 \times -30) - (6 \times 29)] - 1[(7 \times 29) - (-8 \times -47)]$$

$$= 305 \neq 0$$

showing that the system (3.39) with output (3.61) is observable.

EXERCISE 3.19 Return to the system described by (3.35), and suppose that the output is

$$y(k) = -x_1(k) + 3x_2(k)$$

(a) Determine the observability matrix V and show that the system is observable.
(b) If $y(0) = -5$ and $y(1) = 1$, use (3.59) to determine the initial state $x(0)$.

EXERCISE 3.20 Consider the system described by

$$x(k+1) = \begin{bmatrix} 1 & 2 \\ -1 & 4 \end{bmatrix} x(k)$$

$$y(k) = \beta x_1(k) + x_2(k), \quad k = 0, 1, 2, \ldots$$

Find for what values of the parameter β the system is *not* observable.

EXERCISE 3.21 Test the system (3.52) for observability when A is the matrix in Exercise 3.14 and the output is $y(k) = x_2(k)$.

The way we have defined the concepts of controllability and observability shows that they are not connected. Indeed, it's certainly possible for a controllable system to be not observable, or for an observable system to be not controllable, as the following example illustrates.

■ *EXAMPLE 3.12*

Consider the system described by

$$x(k+1) = \begin{bmatrix} 1 & 2 \\ -1 & 4 \end{bmatrix} x(k) + \begin{bmatrix} \alpha \\ 1 \end{bmatrix} u(k) \tag{3.62}$$

$$y(k) = \beta x_1(k) + x_2(k), \quad k = 0, 1, 2, \ldots$$

Suppose $\alpha = 1$ and $\beta = 1$: the controllability matrix is

$$U = \begin{bmatrix} 1 & 3 \\ 1 & 3 \end{bmatrix} \begin{matrix} \\ Ab \end{matrix}$$
$$b$$

and the observability matrix is

$$V = \begin{bmatrix} 1 & 1 \\ 0 & 6 \end{bmatrix} \begin{matrix} c \\ cA \end{matrix}$$

Using (3.34) gives $\det U = 0$, $\det V = 6$, so the system is observable but not controllable. You may have noticed that (3.62) is a combination of the equations in Exercises 3.13 and 3.20, where you were asked to find values of the parameters α and β for which the system was either not controllable or not observable. These 'critical' values of α and β are independent of each other. Thus, for example, if $\alpha \neq 1, 2$ and $\beta = -1$ we have controllability but not observability.

In view of the above example, it's therefore intriguing that *mathematically* the conditions $\det U \neq 0$ for controllability and $\det V \neq 0$ for observability are so similar: the matrix U has columns $b, Ab, A^2b, \ldots, A^{n-1}b$, and the matrix V has rows $c, cA, cA^2, \ldots, cA^{n-1}$. This similarity can be exploited to construct what are called 'dual systems' (see Problem 3.16). These can be used to obtain many useful results involving controllability and observability.

EXERCISE 3.22 Show that the system

$$x(k+1) = \begin{bmatrix} 3 & 1 \\ -1 & 5 \end{bmatrix} x(k)$$

$$y(k) = -x_1(k) + x_2(k)$$

is unobservable. Show also that when

$$x(0) = \begin{bmatrix} 2 \\ 2 \end{bmatrix} \tag{3.63}$$

then the corresponding output is $y(k) = 0$ for all $k \geqslant 0$.

Notice in Exercise 3.22 that if the initial state $x(0)$ is zero then the subsequent output is also zero; hence, there is no way of distinguishing the initial state (3.63) from $x(0) = 0$ by measuring the output. This agrees with our definition of an unobservable system.

To end this section we can consider the *differential* equation model

$$\frac{dx}{dt} = Ax, \quad y = cx \tag{3.64}$$

where $x(t)$, $y(t)$ now depend upon the continuous time variable t. As we saw for the case of controllability in the previous section, the mathematical condition for observability is exactly as before in (3.60), namely that the observability matrix V has a non-zero determinant.

■ *EXAMPLE 3.13*

We looked at the controllability of the electrically heated oven problem in Example 3.10. Suppose it is only possible to measure x_2, the excess of the jacket temperature over the surrounding temperature. We ask whether it is possible to find the oven (excess) temperature x_1 by measuring the output $y = x_2$. We therefore have $c = [0, 1]$, and the matrix A was given in Example 3.10, so we can construct

$$V = \begin{bmatrix} 0 & 1 \\ 10 & -25 \end{bmatrix} \begin{matrix} c \\ cA \end{matrix}$$

and

$$\det V = -10 \neq 0$$

This shows that the system is observable, so the oven interior temperature can indeed be determined merely by measuring the temperature of the jacket.

EXERCISE 3.23 Suppose that for the rabbit–fox population model described in Exercise 3.18 it is only possible to count the *total* number of animals. Is it nevertheless possible to determine the individual numbers of rabbits and of foxes?

EXERCISE 3.24 Return again to the mechanical system described by (3.12), illustrated in Figure 3.4. If it is only possible to measure the displacement x_1 of the mass, can its velocity x_2 be determined?

We'll now show how to obtain an actual expression for the initial state in the differential equation case (3.64), corresponding to (3.59). From (3.64) differentiating y gives

$$\begin{aligned}\frac{dy}{dt} &= c\,\frac{dx}{dt} \\ &= cAx \end{aligned} \tag{3.65}$$

since c is a constant vector. Similarly, differentiating (3.65) gives

$$y^{(2)} = \frac{d^2y}{dt^2} = cA\,\frac{dx}{dt} = cA^2x \tag{3.66}$$

$$\vdots$$

$$y^{(n-1)} = cA^{n-1}x \tag{3.67}$$

where $y^{(i)}$ denotes the ith derivative of y with respect to t. Now set $t=0$ in each of the equations (3.64) to (3.67), giving

$$y(0) = cx(0), \quad y^{(1)}(0) = cAx(0) \tag{3.68}$$

$$y^{(2)}(0) = cA^2x(0), \ldots, y^{(n-1)}(0) = cA^{n-1}x(0)$$

where $y^i(0)$ denotes the value of the ith derivative of y at $t=0$. Writing the n expressions in (3.68) in combined form gives

$$\begin{bmatrix} y(0) \\ y^{(1)}(0) \\ y^{(2)}(0) \\ \vdots \\ y^{(n-1)}(0) \end{bmatrix} = Vx(0)$$

so that provided the system is observable we can write

$$x(0) = V^{-1}\begin{bmatrix} y(0) \\ y^{(1)}(0) \\ \vdots \\ y^{(n-1)}(0) \end{bmatrix} \tag{3.69}$$

This is an explicit expression for $x(0)$ which requires a knowledge of the output y and its first $n-1$ derivatives evaluated at $t=0$.

EXERCISE 3.25 A system modelled by the linear differential equations (3.64) with

$$A=\begin{bmatrix}-1 & -1\\ 2 & -4\end{bmatrix}, \quad c=[1,2]$$

is found to have a scalar output

$$y(t)=-20e^{-3t}+21e^{-2t}$$

Verify that the system is observable, and hence obtain $x(0)$ using (3.69).

3.4 LINEAR FEEDBACK

We introduced the crucial concept of feedback in Section 3.1. In essence, this means that the control applied to a system takes account of the current state of that system – information about the state is 'fed back' to the controller, which reacts appropriately. We shall only consider systems with a single input, in either discrete or continuous form, respectively

$$x(k+1)=Ax(k)+bu(k), \quad k=0,1,2,\ldots \tag{3.70}$$

$$\frac{dx(t)}{dt}=Ax(t)+bu(t) \tag{3.71}$$

where as before x is the $n\times 1$ column vector describing the state, A is an $n\times n$ matrix and b is a constant $n\times 1$ column vector. The idea of *linear feedback* is to make the control a *linear combination* of the states, that is

$$u=f_1x_1+f_2x_2+\cdots+f_nx_n \tag{3.72}$$

or

$$u=fx \tag{3.73}$$

where $f_1, f_2, \ldots, f_n$ are constants, which are the components of the row *feedback vector* f. If we apply (3.73) to (3.70) and (3.71) we obtain

$$\begin{aligned} x(k+1)&=(A+bf)x(k)\\ \frac{dx(t)}{dt}&=(A+bf)x(t)\end{aligned} \tag{3.74}$$

which are called *closed loop* systems. In either case the matrix of the system is

$$\mathcal{A}=A+bf \tag{3.75}$$

and is called the *closed loop matrix*. A key *theorem* discovered around 1960 states that if the system (either version) in (3.74) is controllable then it is always possible

to determine a vector f such that the eigenvalues of the matrix $\mathcal{A}$ in (3.75) can be made to take *any* preselected set of n values (subject only to the condition that any complex values occur in conjugate pairs).

You may feel rather intimidated by the apparent complexity of this theorem, so let's explore in some detail what it means. First, we substitute the expression (3.73) for the feedback control into the equations (3.70) and (3.71). The complete system which results is then given by (3.74), and is called 'closed loop' for the reason given in Section 3.1 (see Figure 3.2). The systems in (3.74) for the discrete and continuous time cases are respectively

$$x(k+1) = \mathcal{A}x(k) \tag{3.76}$$

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = \mathcal{A}x(t) \tag{3.77}$$

where $\mathcal{A}$ is the matrix in (3.75). We saw how to solve difference equations of the type (3.76) in Chapter 1, Section 1.4. It was found that

$$x(k) = \mathcal{A}^k x(0) \tag{3.78}$$

Furthermore, an expression was given in (1.106) for $\mathcal{A}^k$ which depended upon the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ of $\mathcal{A}$. You may remember that the eigenvalues are defined as the roots of the nth-degree polynomial equation

$$\det(\lambda I - \mathcal{A}) = 0 \tag{3.79}$$

called the characteristic equation of $\mathcal{A}$. Never mind the details of how these eigenvalues are calculated. What matters is that the solution $x(k)$ given in (3.78) of the closed loop system depends crucially on these eigenvalues: if we can control them, then we can control the behaviour of $x(k)$ – which is, after all, our prime objective. Our theorem therefore adds precision to the property of controllability, which stated that a control satisfying required objectives *could* be found – the theorem tells us just what can be achieved using linear feedback. There is only one minor restriction, caused by a law of algebra: since the polynomial equation (3.79) has *real* coefficients its roots either are real or occur in complex conjugate pairs $\alpha + \mathrm{i}\beta$, $\alpha - \mathrm{i}\beta$. Apart from this, we can make the eigenvalues of $\mathcal{A}$ equal to any set of values we choose by selecting an appropriate feedback vector f. For this reason the theorem is often called the *eigenvalue assignment* theorem.

Although not covered in this book, similar remarks concerning the solution of the differential equation (3.77) apply: the solution again depends upon the eigenvalues of $\mathcal{A}$ (actually involving terms in $\mathrm{e}^{\lambda_i t}$ rather than $(\lambda_i)^k$), and making the λ equal to a predetermined set of values largely determines the way the system behaves.

In either case, if the system is not controllable then assignment of arbitrary eigenvalues to $\mathcal{A}$ is not possible.

■ EXAMPLE 3.14

We wish to find linear feedback such that when applied to the system

$$x(k+1)=\begin{bmatrix}1 & -3\\ 4 & 2\end{bmatrix}x(k)+\begin{bmatrix}1\\ 2\end{bmatrix}u \tag{3.80}$$

the resulting closed loop system has eigenvalues -1, -2.

You should check that the system is controllable. The closed loop matrix in (3.75) is

$$\begin{aligned}\mathcal{A} &= A+bf\\ &=\begin{bmatrix}1 & -3\\ 4 & 2\end{bmatrix}+\begin{bmatrix}1\\ 2\end{bmatrix}[f_1 \;\; f_2]\\ &=\begin{bmatrix}1 & -3\\ 4 & 2\end{bmatrix}+\begin{bmatrix}f_1 & f_2\\ 2f_1 & 2f_2\end{bmatrix}\\ &=\begin{bmatrix}1+f_1 & -3+f_2\\ 4+2f_1 & 2+2f_2\end{bmatrix}\end{aligned} \tag{3.81}$$

Notice that the product of the column vector b and the row vector f produces a 2×2 matrix, using the rule set out in Chapter 1, Section 1.4. In general the product bf is equal to an $n\times n$ matrix with $b_i f_j$ being the element in row i, column j. The eigenvalues of $\mathcal{A}$ are given by solving (3.79), which here is

$$\det(\lambda I-\mathcal{A})=\det\begin{bmatrix}\lambda-1-f_1 & 3-f_2\\ -4-2f_1 & \lambda-2-2f_2\end{bmatrix}=0$$

Using the formula (3.34) we can evaluate this determinant as

$$\begin{aligned}\det(\lambda I-\mathcal{A}) &= (\lambda-1-f_1)(\lambda-2-2f_2)-(3-f_2)(-4-2f_1)\\ &= \lambda^2-\lambda(f_1+2f_2+3)+(8f_1-2f_2+14)\end{aligned}$$

For $\mathcal{A}$ to have eigenvalues -1, -2 its characteristic polynomial must be

$$(\lambda+1)(\lambda+2)=\lambda^2+3\lambda+2$$

Comparing the coefficients of these two quadratics produces the equations

$$-f_1-2f_2-3=3$$
$$8f_1-2f_2+14=2$$

which are easily found to have the solution $f_1=-2$, $f_2=-2$. The required feedback (3.72) is therefore

$$u=-2x_1-2x_2$$

EXERCISE 3.26 If

$$A=\begin{bmatrix}1 & 3\\ -1 & 4\end{bmatrix},\quad b=\begin{bmatrix}1\\ 3\end{bmatrix}$$

determine a feedback vector f such that the eigenvalues of $A+bf$ are -2, 3.

EXERCISE 3.27 If

$$A=\begin{bmatrix}0 & 1 & 0\\ 0 & 0 & 1\\ 6 & -11 & 6\end{bmatrix},\quad b=\begin{bmatrix}0\\0\\1\end{bmatrix}$$

determine f such that $A+bf$ has eigenvalues $-1,\ -2\pm 3\mathrm{i}$.

EXERCISE 3.28 If

$$A=\begin{bmatrix}2 & 0\\ 0 & 2\end{bmatrix},\quad b=\begin{bmatrix}1\\3\end{bmatrix},\quad f=[f_1\ f_2]$$

show that the system is not controllable. Show also that $A+bf$ has one eigenvalue equal to 2 whatever the values of f_1 and f_2, so it is *not* possible to assign arbitrary eigenvalues to the closed loop matrix.

EXERCISE 3.29 Show that the system described by

$$\frac{dx}{dt}=\begin{bmatrix}-2 & 0 & 3\\ 0 & -3 & 0\\ 1 & 0 & -4\end{bmatrix}x+\begin{bmatrix}1\\1\\-1\end{bmatrix}u$$

is not controllable. Show also that with linear feedback of the form $u=f_1x_1+f_3x_3$ the closed loop system has two fixed eigenvalues, one of which is equal to -3. Determine the second fixed eigenvalue and values of f_1 and f_3 such that the third eigenvalue of $A+bf$ is equal to -4.

When $n>2$ the method used in Example 3.14 is not satisfactory in general, and a number of other techniques have been devised. One relies on transforming the matrix A into 'companion form' and the vector b into a special form (see Problem 3.14). Details are outside the scope of this book.

■ *EXAMPLE 3.15*

Let's return again to the rabbit–fox population model described in Exercise 3.18. Suppose that the values of the constants a, b, c, d in (3.47) and (3.48) are such that

$$\frac{dx_1}{dt}=2x_1-3x_2,\quad \frac{dx_2}{dt}=2x_1-x_2 \tag{3.82}$$

As indicated earlier, one way to control a 'population explosion' is to introduce linear feedback in the form of a disease which affects rabbits but not foxes. This means that the rate of growth of the rabbit population is reduced by an amount fx_1, where f is a positive parameter. In other words, the first equation in (3.82) becomes

$$\frac{dx_1}{dt}=2x_1-3x_2-fx_1$$

Combining this with the second equation in (3.82) shows that the closed loop matrix is

$$\mathcal{A} = \begin{bmatrix} (2-f) & -3 \\ 2 & -1 \end{bmatrix}$$

The eigenvalues of this matrix are given by

$$\begin{aligned} 0 &= \det(\lambda I - \mathcal{A}) \\ &= \begin{bmatrix} \lambda - 2 + f & 3 \\ -2 & \lambda + 1 \end{bmatrix} \\ &= \lambda^2 + \lambda(f-1) + f + 4 \end{aligned} \tag{3.83}$$

As we have mentioned, the solution $x_1(t)$, $x_2(t)$ of the closed loop system involves terms $e^{\lambda_1 t}$, $e^{\lambda_2 t}$ where λ_1 and λ_2 are the roots of (3.83). If either of the roots has a positive real part then the exponential term gets larger and larger as t increases – this is a 'population explosion'. If both the roots have a negative real part then $e^{\lambda t} \to 0$ as $t \to \infty$, so the population declines to zero after a sufficiently long time. If λ_1 and λ_2 are purely imaginary, that is $\lambda_i = iw$, $\lambda_2 = -iw$ where w is real, then as we saw in Section 1.2, Chapter 1,

$$e^{\lambda t} = \cos\ wt \pm i \sin\ wt$$

so the population oscillates in size but remains finite. The roots of (3.83) have the form

$$\frac{1 - f \pm \sqrt{[(f-1)^2 - 4(f+4)]}}{2}$$

You should be able to see that the real parts of the roots are positive when $0 < f < 1$ and negative when $f > 1$; when $f = 1$ the roots are purely imaginary. We therefore conclude that the smallest value of the feedback parameter which will prevent the numbers of rabbits and foxes from growing ever larger is $f = 1$.

EXERCISE 3.30 A system has

$$A = \begin{bmatrix} 1 & -1 \\ 2 & 3 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

and linear feedback $u = f_1 x_1 + f_2 x_2$ is applied. Determine the conditions to be satisfied by f_1 and f_2 so that the eigenvalues of the closed loop matrix are purely imaginary. (Hint: you'll need to find the form a quadratic equation must take in order to have purely imaginary roots iw and $-iw$, where w is real.)

3.5 MULTIPLE CONTROLS AND OUTPUTS

We now look at the more complicated situation where there are several control variables. This section is rather more difficult than the rest of the chapter, and can be omitted on a first reading of the book. Let's begin with the discrete equation

$$x(k+1) = Ax(k) + Bu(k), \quad k = 0, 1, 2, \ldots \tag{3.84}$$

where now $u(k)$ is a column vector consisting of m control variables $u_1(k)$, $u_2(k), \ldots, u_m(k)$ with $m > 1$. If there are n state variables $x_1(k), x_2(k), \ldots, x_n(k)$ then in (3.84) the matrix A is $n \times n$ and B is $n \times m$. In practice it will always be the case that $m \leqslant n$, and in fact usually $m < n$ so that B is a rectangular matrix. An example of (3.84) with $n = 3$ and $m = 2$ was given in Problem 1.26, Chapter 1, as the model of a trout fish farm. Suppose we investigate the controllability problem using the same sort of argument that we applied in Example 3.8 in the previous section. Without going into details it turns out that the *controllability matrix* now becomes

$$U = [B, AB, A^2B, \ldots, A^{n-1}B] \tag{3.85}$$

which is exactly the same as (3.41) except that the column vector b is replaced by the $n \times m$ matrix B. Notice that U in (3.85) has n rows as before, but now has mn columns. The controllability condition therefore cannot now be that U is non-singular (or equivalently $\det U \neq 0$) since the concepts of non-singularity and determinant only apply when U is square. Instead, the controllability condition is

$$\text{rank } U = n \tag{3.86}$$

What does this mean? We must now spend some time explaining the idea of the rank of a matrix, and then look at a method of computing the rank. Before doing so, let's work out U in a simple case.

■ *EXAMPLE 3.16*

Let the matrices in (3.84) be

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \tag{3.87}$$

so that $n = 3$ and $m = 2$. The product AB is found using the rule explained in Section 1.4 of Chapter 1 (see equation (1.64)), giving

$$AB = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 0 \end{bmatrix}$$

and

$$A^2B = A(AB)$$

$$= \begin{bmatrix} 2 & 1 \\ 1 & 0 \\ 1 & 0 \end{bmatrix}$$

so that the controllability matrix (3.85) is

$$U = [B, AB, A^2B]$$

$$= \begin{bmatrix} 0 & 1 & 1 & 1 & 2 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \end{bmatrix} \qquad (3.88)$$

$$c_1 \quad c_2 \quad c_3 \quad c_4 \quad c_5 \quad c_6$$

The condition for controllability is that the rank (yet to be defined!) of the matrix in (3.88) must be three.

For our present purpose we can define the *rank* of a matrix as the largest number of *independent* columns which it possesses. A set of columns of a matrix is called *independent* if no column in this set can be expressed as a linear combination of any of the other columns in the set. For example, the first and second columns c_1 and c_2 in (3.88) are independent of each other since there is no way that we can express c_2 as a multiple of c_1. Remember that a linear combination of columns $c_1, c_2, c_3, \ldots$ simply means the expression $\alpha_1 c_1 + \alpha_2 c_2 + \alpha_3 c_3 + \cdots$, where the α are constants, not all zero. From (3.88) we see that

$$\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

$$c_3 \qquad c_1 \qquad c_2$$

so c_3 is *not* independent – it is said to be *dependent* upon c_1 and c_2 since it can be expressed as a linear combination of them. Similarly you can see that the other columns in (3.88) satisfy

$$c_4 = c_2, \quad c_5 = c_1 + 2c_2, \quad c_6 = c_2$$

so all the other columns of (3.88) can be expressed in terms of c_1 and c_2. Since the largest number of independent columns of U in (3.88) is therefore two, its rank is *two*. Hence the system with matrices A and B in (3.87) is *not* controllable, since it does not have rank 3.

■ *EXAMPLE 3.17*

The matrix

$$M = \begin{bmatrix} 1 & 4 & 1 & -1 & 4 \\ 2 & 8 & 2 & -2 & 8 \\ 3 & 12 & 0 & -6 & 9 \end{bmatrix} \qquad (3.89)$$

$$c_1 \quad c_2 \quad c_3 \quad c_4 \quad c_5$$

has rank 2, since

$$c_2 = 4c_1, \quad c_4 = c_3 - 2c_1, \quad c_5 = 3c_1 + c_3$$

but c_1 and c_3 are independent – we cannot express c_3 in the form αc_1 for any value of α. Hence there are *two* independent columns of M, so M has rank 2.

EXERCISE 3.31 For what values of k will the matrix

$$M = \begin{bmatrix} 1 & 2 & 3 & 2 \\ 2 & 4 & 6 & 4 \\ 5 & 10 & 15 & k \end{bmatrix}$$

have rank equal to (a) 1, (b) 2, (c) 3?

Some relevant facts about rank are:

(i) A matrix has rank 0 only if *all* its elements are zero.

(ii) The rank of a matrix cannot be bigger than the *smaller* of its two dimensions. For example, a 3×4 matrix cannot have rank bigger than three.

(iii) In view of (ii), the $n \times mn$ controllability matrix U in (3.85) cannot have rank bigger than n. Hence the controllability condition requires that rank U has its maximum possible value.

We now need to describe a method for computing the rank of a matrix. The method is called *gaussian elimination* after the German supermathematician Gauss (he worked in the first half of the nineteenth century).

We shan't go into details as to why the method works – these belong in a book on matrix algebra.

(i) We begin by considering a square (*upper*) *triangular* matrix, that is all the elements *below* the *principal diagonal* (northwest to southeast) are zero. For example,

$$M_1 = \begin{bmatrix} 1 & 3 & 4 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.90}$$

is a 3×3 triangular matrix. The elements on the diagonal are called the *pivots*, and the rank is simply equal to the number of non-zero pivots. Thus M_1 in (3.90) has rank 3 (the pivots are 1, 2, 1) and

$$M_2 = \begin{bmatrix} 1 & 3 & 4 & 1 \\ 0 & 2 & 5 & -1 \\ 0 & 0 & 3 & 2 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{3.91}$$

has rank 3 since there are three non-zero points (1, 2, 3). In each case the diagonal has been indicated by a dotted line. Notice that the last row of M_2 in (3.91) consists of all zeros.

(ii) In the same way, the rank of a rectangular matrix which *begins* with a triangular block is *also* equal to the number of non-zero pivots, as the following example illustrates:

$$M_3 = \left[\begin{array}{ccc|cc} 1 & 3 & 4 & 2 & 4 \\ 0 & 2 & 0 & 5 & 6 \\ 0 & 0 & 1 & 1 & 0 \end{array}\right], \quad \text{rank } M_3 = 3$$

$$\leftarrow M_1 \rightarrow$$

The triangular block to the left of the dashed line is in fact the matrix M_1 in (3.90).

This property also holds if there are rows of zeros in the triangular block, *provided these continue right along the matrix*, as illustrated by

$$M_4 = \left[\begin{array}{cccc|ccc} 1 & 3 & 4 & 1 & -1 & 2 & 7 \\ 0 & 2 & 5 & -1 & 4 & 9 & 3 \\ 0 & 0 & 3 & 2 & 1 & -2 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right], \quad \text{rank } M_4 = 3$$

$$\longleftarrow M_2 \longrightarrow$$

Another example is

$$M_5 = \left[\begin{array}{cccc|ccc} 1 & 3 & 4 & 1 & 0 & 2 & 1 \\ 0 & 2 & 5 & -1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

where rank $M_5 = 2$ because of the non-zero pivots 1, 2. However, for the matrix

$$M_6 = \left[\begin{array}{cccc|ccc} 1 & 3 & 4 & 1 & 0 & 2 & 1 \\ 0 & 2 & 5 & -1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 4 & 0 \end{array}\right] \tag{3.92}$$

$$c_1 \quad c_2 \quad c_3 \quad c_4 \quad c_5 \quad c_6 \quad c_7$$

we cannot say that rank $M_6 = 2$, because the bottom two rows of zeros in the triangular block (the first four columns of M_6) have some non-zero elements along the remainder of the rows in the last three columns of M_6.

(iii) To handle a matrix like M_6 in (3.92), we need the fact that swopping around columns of a matrix *does not alter its rank*. In (3.92) if we interchange the third and seventh columns, and the fourth and sixth

columns, we get

$$\left[\begin{array}{ccccc|ccc} 1 & 3 & 1 & 2 & & 0 & 1 & 4 \\ 0 & 2 & 0 & 1 & & 1 & -1 & 5 \\ 0 & 0 & -1 & 0 & & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & & 0 & 0 & 0 \end{array}\right]$$

$$c_1 \quad c_2 \quad c_7 \quad c_6 \quad c_5 \quad c_4 \quad c_3$$

Since this now contains *four* non-zero pivots (1, 2, −1, 4) it follows that rank $M_6 = 4$. It is convenient to denote the column swops by $c_3 \leftrightarrow c_7$ and $c_4 \leftrightarrow c_6$.

(iv) We can use the following *elementary row operations* to reduce a rectangular matrix to the form described in (ii):

Interchange any two rows of the matrix: $r_i \leftrightarrow r_j$ denotes swopping round the ith and jth rows.

Add an arbitrary multiple of any row to any other row: $r_i + pr_j$ denotes adding p times row j to row i, where p is any positive or negative number.

It's best to consider some examples to see how the method works in practice.

■ *EXAMPLE 3.18*

(a) Consider the controllability matrix U in (3.88). Our objective is to produce a triangular block with a maximum possible number of non-zero pivots. This is done by applying appropriate column swops and/or elementary row operations. In order to begin, we must get a non-zero pivot in the (1, 1) position of the matrix, so we interchange columns 1 and 2 in (3.88) to get

$$\begin{bmatrix} 1 & 0 & 1 & 1 & 2 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \end{bmatrix}$$

We now subtract row 2 from row 3 (i.e. $r_3 - r_2$) to get

$$\left[\begin{array}{ccc|ccc} 1 & 0 & 1 & 1 & 2 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

and this has the required form – no further reduction is possible. Since the triangular block to the left of the dashed line has two non-zero pivots, the rank is 2, agreeing with what we found earlier.

(b) Consider the matrix M in (3.89) and apply operations to it as indicated, so as to obtain the first column of the triangular block, with zeros below the first pivot:

$$M \xrightarrow[r_3 - 3r_1]{r_2 - 2r_1} \begin{bmatrix} 1 & 4 & 1 & -1 & 4 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -3 & -3 & -3 \end{bmatrix}$$

The long arrow indicates what happens to M when twice row 1 is subtracted from row 2, and three times row 1 from row 3. We haven't finished, because we can still get another non-zero pivot by moving any of the elements -3 to the (2, 2) position. For example, swop rows 2 and 3, then swop columns 2 and 3:

$$\xrightarrow{r_2 \leftrightarrow r_3} \begin{bmatrix} 1 & 4 & 1 & -1 & 4 \\ 0 & 0 & -3 & -3 & -3 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\xrightarrow{c_2 \leftrightarrow c_3} \left[\begin{array}{ccc|cc} 1 & 1 & 4 & -1 & 4 \\ 0 & -3 & 0 & -3 & -3 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

The required form has now been obtained, showing that the rank is 2 because of the two non-zero pivots 1, -3.

(c) The following example is self-explanatory:

$$M = \begin{bmatrix} 3 & -2 & 0 & 1 \\ -1 & 1 & 2 & 2 \\ 1 & 0 & 4 & 5 \end{bmatrix} \xrightarrow[r_3 - \frac{1}{3}r_1]{r_2 + \frac{1}{3}r_1} \begin{bmatrix} 3 & -2 & 0 & 1 \\ 0 & \frac{1}{3} & 2 & \frac{7}{3} \\ 0 & \frac{2}{3} & 4 & \frac{14}{3} \end{bmatrix}$$

$$\xrightarrow{r_3 - 2r_2} \left[\begin{array}{ccc|c} 3 & -2 & 0 & 1 \\ 0 & \frac{1}{3} & 2 & \frac{7}{3} \\ 0 & 0 & 0 & 0 \end{array}\right]$$

The pivots 3, 1/3, 0 show that rank $M = 2$.

To summarize the procedure:

(i) If necessary, apply column and/or row swops to get a non-zero pivot in the (1, 1) position.

(ii) Apply elementary row operations to reduce all the elements in the first column below the first pivot to zero.

(iii) Repeat (i) if necessary for a non-zero second pivot, in the (2, 2) position; repeat (ii) so that all the elements in the second column below the pivot are zero.

(iv) Continue like this until the required form containing a triangular block is obtained.

EXERCISE 3.32 Determine the rank of the following matrices:

(a)

$$\begin{bmatrix} 1 & 3 & 2 & 3 \\ 2 & 6 & 5 & 7 \\ 3 & 10 & 7 & 11 \end{bmatrix}$$

(b)

$$\begin{bmatrix} 3 & 6 & -5 & 10 \\ 2 & 4 & -4 & 7 \\ 2 & 4 & -6 & 8 \end{bmatrix}$$

(c)

$$\begin{bmatrix} 1 & 0 & 1 & -1 & 0 & 1 \\ 1 & 1 & 2 & 3 & 4 & 3 \\ 2 & 1 & 3 & 2 & 4 & 4 \\ 1 & -2 & 1 & -9 & -8 & -3 \\ 7 & 5 & 12 & 13 & 20 & 17 \end{bmatrix}$$

EXERCISE 3.33 Compute the controllability matrix U in (3.85) when

$$A = \begin{bmatrix} 1 & 2 & 0 \\ 3 & 4 & 1 \\ -1 & 2 & -5 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 3 & 2 \end{bmatrix} \tag{3.93}$$

Determine rank U, and hence decide whether the system (3.84) is controllable in this case.

EXERCISE 3.34 Show that the system (3.84) is controllable when

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 3 & 0 & 0 & 2 \\ 0 & 0 & 0 & 1 \\ 0 & -2 & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}$$

If one of the control variables ceases to operate, that is *either* $u_1 = 0$ *or* $u_2 = 0$, test whether the system remains controllable in each case. Notice that in each of these cases B reduces to a column vector only.

EXERCISE 3.35 If

$$A = \begin{bmatrix} 2 & (a-3) \\ 0 & 2 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 1 \\ 0 & (a^2 - a) \end{bmatrix}$$

determine for what values of the parameter a the system (3.84) is *not* controllable. Investigate the situation if the first control variable ceases to operate, that is $u_1 = 0$ (compare with Exercise 3.17).

Before going on to consider observability for the case of several outputs, it's useful to note what happens when gaussian elimination is applied to a *square* $n \times n$ matrix M. In this case the rank is still equal to the number of non-zero pivots after the matrix has been reduced to triangular form. If any pivot is zero then $\det M = 0$; if all the pivots are non-zero then

$$\det M = (-1)^t \times \text{product of the pivots}$$

where t is the total number of row and column interchanges (if any).

■ EXAMPLE 3.19

(a) The determinant of the triangular matrix M_1 in (3.90) is $1 \times 2 \times 1 = 2$. The determinant of M_2 in (3.91) is zero, since there is a zero pivot.

(b) The determinant of *any* triangular matrix is just equal to the product of the elements on the principal diagonal, for example

$$\det \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{bmatrix} = a_{11}a_{22}a_{33}$$

(c)

$$\begin{bmatrix} 3 & -2 & 0 \\ -1 & 1 & 2 \\ 1 & 0 & 6 \end{bmatrix} \xrightarrow[r_3 - \frac{1}{3}r_1]{r_2 + \frac{1}{3}r_1} \begin{bmatrix} 3 & -2 & 0 \\ 0 & \frac{1}{3} & 2 \\ 0 & \frac{2}{3} & 6 \end{bmatrix}$$

$$\xrightarrow{r_3 - 2r_1} \begin{bmatrix} 3 & -2 & 0 \\ 0 & \frac{1}{3} & 2 \\ 0 & 0 & 2 \end{bmatrix}$$

The determinant of the original matrix is therefore $3 \times \frac{1}{3} \times 2 = 2$.

(d) Consider the controllability matrix U in Example 3.9:

$$U = \begin{bmatrix} 1 & 3 & -18 \\ 0 & 7 & -19 \\ 1 & -7 & 41 \end{bmatrix} \xrightarrow{r_3 - r_1} \begin{bmatrix} 1 & 3 & -18 \\ 0 & 7 & -19 \\ 0 & -10 & 59 \end{bmatrix}$$

$$\xrightarrow{r_3 + \frac{10}{7}r_2} \begin{bmatrix} 1 & 3 & -18 \\ 0 & 7 & -19 \\ 0 & 0 & \frac{223}{7} \end{bmatrix}$$

so det $U = 1 \times 7 \times (223/7) = 223$, agreeing with what was found earlier.

For determinants having numerical elements, evaluation using gaussian elimination is always preferable to using expansion formulae like that in (1.86). It is also worth mentioning that an extension of gaussian elimination provides a good way of computing the inverse of a matrix.

EXERCISE 3.36 Use gaussian elimination to evaluate det V, where V is the observability matrix in Example 3.11.

EXERCISE 3.37 Use gaussian elimination to evaluate

$$\det \begin{bmatrix} 1 & 2 & 3 & 4 \\ -1 & 1 & 2 & 3 \\ 1 & -1 & 1 & 2 \\ 2 & -2 & 2 & 11 \end{bmatrix}$$

EXERCISE 3.38 Show that the system with matrices

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ 2 \\ 0 \\ 1 \end{bmatrix}$$

is not controllable.

To complete our discussion on controllability with multiple controls, we note that the condition (3.86) still applies to the differential equation description (3.7), which we repeat for convenience:

$$\frac{dx}{dt} = Ax + Bu \tag{3.94}$$

where u is again an $m \times 1$ column vector and B is $n \times m$. The expression for U in (3.85) is also unchanged.

Let's now move on to observability when there are multiple outputs, so the output is a column vector y with r components $y_1, \ldots, y_r$. As before we assume linearity, meaning that each output variable can be expressed as a linear combination of the states. We can therefore write

$$y = Cx \tag{3.95}$$

where C is an $r \times n$ matrix and x is the state vector. Equation (3.95) can apply either to the difference equation model (3.84) or to the differential equation model (3.94).

Notice that if C is square and non-singular then we immediately obtain from (3.95)

$$x = C^{-1}y$$

showing that the state can certainly be determined from the output, so by definition the system is observable. Otherwise, we have $r < n$ and it turns out that the condition for observability is

$$\text{rank } V = n \tag{3.96}$$

where V is the *observability matrix*

$$V = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{n-1} \end{bmatrix} \tag{3.97}$$

Notice that V has n columns and rn rows. When $r = 1$ then (3.97) reduces to the previous expression in (3.57) for the single output case. In order to compute the rank of V in (3.97) we use the fact that if the rows and columns of a matrix are

interchanged this does not affect the rank – this is called *transposing* a matrix, and is denoted by V^T (see Problem 3.16). The method of gaussian elimination can then be applied to V^T, exactly as was done for U.

■ EXAMPLE 3.20

Let the matrices in (3.84) (or (3.94)) and (3.95) be

$$A=\begin{bmatrix}1&1&0\\0&0&1\\0&1&0\end{bmatrix},\quad C=\begin{bmatrix}1&1&0\\1&0&1\end{bmatrix}$$

so that $n=3$ and $r=2$. Using our multiplication rule we get

$$CA=\begin{bmatrix}1&1&1\\1&2&0\end{bmatrix}$$

$$CA^2=(CA)A=\begin{bmatrix}1&2&1\\1&1&2\end{bmatrix}$$

so that

$$V=\begin{bmatrix}1&1&0\\1&0&1\\1&1&1\\1&2&0\\1&2&1\\1&1&2\end{bmatrix} \tag{3.98}$$

To obtain V^T, we write each row of (3.98) as a column to obtain

$$V^T=\begin{bmatrix}1&1&1&1&1&1\\1&0&1&2&2&1\\0&1&1&0&1&2\end{bmatrix} \tag{3.99}$$

The first row of (3.98) becomes the first column in (3.99), the second row of (3.98) is the second column in (3.99), and so on. To compute the rank of (3.99) we then apply the gaussian elimination procedure as follows:

$$V^T \xrightarrow{r_2-r_1} \begin{bmatrix}1&1&1&1&1&1\\0&-1&0&1&1&0\\0&1&1&0&1&2\end{bmatrix}$$

$$\xrightarrow{r_3+r_2} \left[\begin{array}{ccc:ccc}1&1&1&1&1&1\\0&-1&0&1&1&0\\0&0&1&1&2&2\end{array}\right]$$

There are three non-zero pivots in the triangular block to the left of the dashed line, showing rank $V=3$ and hence the system is observable.

EXERCISE 3.39 Compute the observability matrix V in (3.97) when A is the matrix in (3.93) and

$$C = \begin{bmatrix} 1 & 0 & 2 \\ 2 & 1 & 0 \end{bmatrix}$$

and hence test for observability.

To close this chapter, it's interesting to note that the eigenvalue assignment theorem still holds when there are multiple controls.

Provided the system is controllable it's always possible to find an $m \times n$ constant matrix F such that the closed loop matrix $A + BF$ has preassigned eigenvalues (subject only to the same condition as before, that any complex eigenvalues occur in complex conjugate pairs). The linear feedback now has the form $u = Fx$, meaning that each control variable is a linear combination of the states, that is

$$u_i = f_{i1}x_1 + f_{i2}x_2 + \cdots + f_{in}x_n, \quad i = 1, 2, \ldots, m$$

where f_{ij} is the element in row i, column j of F. Actual determination of such matrices F is well outside the scope of this book.

EXERCISE 3.40 Verify that when

$$A = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}, \quad B = \begin{bmatrix} 2 & 1 \\ 1 & 0 \end{bmatrix}$$

then the linear feedback

$$u_1 = -3x_1 - x_2, \quad u_2 = 3x_1 + x_2$$

produces a closed loop matrix with eigenvalues -3 and -4.

EXERCISE 3.41 Show that if the matrix B in (3.84) is square and non-singular then

$$u(k) = B^{-1}[x(k+1) - Ax(k)]$$

PROBLEMS

3.1 Consider equation (3.20) which describes the motion of a damped mass–spring system shown in Figure 3.7 with $u = 0$, namely

$$m\frac{d^2x}{dt^2} + p\frac{dx}{dt} + kx = 0$$

with $m > 0$, $p > 0$, $k > 0$. Here x is the displacement from rest, and is given by

$$x(t) = \begin{cases} \alpha e^{\lambda_1 t} + \beta e^{\lambda_2 t}, & \lambda_1 \neq \lambda_2 \\ e^{\lambda_1 t}(\gamma + \delta t), & \lambda_1 = \lambda_2 \end{cases}$$

where α, β, γ, δ are arbitrary constants and λ_1, λ_2 are the roots of the quadratic equation

$$m\lambda^2 + p\lambda + k = 0$$

Consider each of the three possibilities for the roots: both real and distinct; both complex; both real and equal; and show that in all cases $x(t) \to 0$ as $t \to \infty$.

3.2 Test for controllability the system described in

(a) Exercise 3.6, equation (3.22)
(b) Exercise 3.7
(c) Exercise 3.9
(d) Exercise 3.10
(e) Exercise 3.11, using slaughter.

3.3 Suppose that in the redwood forest model described in Exercise 1.34 it is feasible only to count *all* the trees in each 50 year period. Can the number of trees in each of the three age groups be determined?

3.4 Consider the cattle ranching model described in Problem 1.23. If it is practicable only to count the total number of cattle, is it possible to determine the number of animals in each age group?

3.5 It was seen in the buffalo population model in Problem 1.28 that under the assumed conditions the numbers of animals would grow by 6.3% per year.

Suppose that (as in Exercise 3.11) a linear feedback slaughtering policy had been adopted, killing for food pF_{k+1} adult females in year $k+1$, so that the first equation in Problem 1.28 is replaced by

$$F_{k+2} = 0.95F_{k+1} + 0.12F_k - pF_{k+1}$$

Use the z-transform (see Section 1.3, Chapter 1) to show that the total population would have remained constant even if 7% of adult females were killed each year.

3.6 The equations describing the vertical motion of a hot air balloon can be expressed as

$$\frac{dT}{dt} = -T + u$$

$$\frac{dv}{dt} = -\tfrac{1}{2}v + T + \tfrac{1}{2}w$$

$$\frac{dh}{dt} = v$$

where the vertical speed is v, the change from equilibrium altitude is h, the change in the temperature of the air in the balloon from the equilibrium temperature is T, and w is the constant vertical wind speed. The control variable u is proportional to the change in heat added to the air in the balloon. Take as state variables

$$x_1 = T, \quad x_2 = v, \quad x_3 = h, \quad x_4 = w$$

and write the equations in the matrix form (3.7). Show that the system is not controllable. Verify that linear feedback

$$u = -\tfrac{1}{4}v - \tfrac{1}{8}h$$

produces a closed loop system with eigenvalues $0, -\frac{1}{2}, -\frac{1}{2}, -\frac{1}{2}$.

3.7 The motion of a helicopter hovering in still air is described by the equations

$$\frac{d^2\theta}{dt^2} + a\frac{d\theta}{dt} = bu$$

$$\frac{d^2s}{dt^2} + c\frac{d\theta}{dt} - d\theta = du$$

where θ is the pitch angle of the fuselage, s is the horizontal distance of the centre of mass of the helicopter from the hover point and a, b, c and d are constants. The control variable u is the tilt angle of the rotor thrust with respect to the fuselage. Take

$$x_1 = \theta, \quad x_2 = s, \quad x_3 = \frac{d\theta}{dt}, \quad x_4 = \frac{ds}{dt}$$

as the state variables and write the equations in the matrix form (3.7).

If $b = 1$, $d = 2$ and $c = 2a$ show that the system is controllable irrespective of the value of a.

3.8 Consider again the mechanical system shown in Figure 3.6. If in addition dampers are connected between each of the masses and the fixed support then the equations (3.19) are replaced by

$$\frac{dx}{dt} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -6 & 3 & -4 & 0 \\ \frac{3}{4} & -(3+k)/4 & 0 & -6 \end{bmatrix} x + \begin{bmatrix} 0 \\ 0 \\ 0 \\ \frac{1}{4} \end{bmatrix} u$$

It is possible to measure only the relative displacement $x_2 - x_1$. Construct the observability matrix. Reduce it to triangular form by gaussian elimination, and hence determine for what values of k the system is *not* observable.

3.9 Determine whether the trout fish farm model described in Problem 1.26 is controllable.

3.10 If

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 6 & -11 & 6 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}$$

verify that linear feedback

$$u_1 = -3x_1 + 4x_2 - x_3, \quad u_2 = -3x_1 + 4x_2 - x_3$$

produces a closed loop system with eigenvalues 1, 1, 3.

3.11 Consider the model of the blue whale population described in Example 1.20. The matrix A in (1.57) has an eigenvalue greater than one, so the population continues to increase in size.

(a) Is it possible to prevent this increase by culling a proportion of females under 4 years old? (Notice that this affects $x_2(k+1)$ only.)

(b) If only the total number of females under 8 years old can be counted, can the numbers of females in each age group be determined?

3.12 Two rods of unit length swing from a fixed support and are connected to each other by a spring as shown in Figure 3.10.

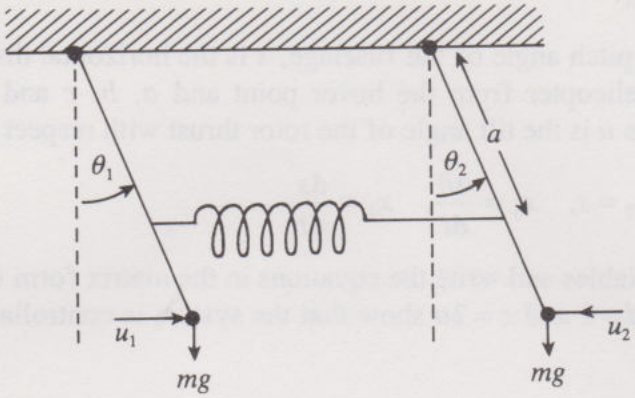


Figure 3.10

Masses m are fixed at the ends of the rods (whose weights can be neglected) and for small oscillations about the vertical the equations of motion are

$$m\frac{dx_3}{dt} = -mgx_1 + u_1 - u_2$$

$$m\frac{dx_4}{dt} = -mgx_2 - 2ka^2x_2 + u_1 + u_2$$

where g is the gravitational constant, u_1 and u_2 are applied control forces, k is the spring constant, $x_1 = \theta_1 + \theta_2$, $x_2 = \theta_1 - \theta_2$ and

$$x_3 = \frac{dx_1}{dt}, \quad x_4 = \frac{dx_2}{dt}$$

Write the equations in the matrix form (3.7). Show that the system is controllable. Show also that if the two forces are equal, that is $u_1 = u_2 = u$, then the system is *not* controllable.

3.13 A certain mechanical system consisting of two masses m_1 and m_2 connected by springs and dampers is described by the equations

$$\frac{dx}{dt} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -k_1/m_1 & k_1/m_1 & -d_1/m_1 & d_1/m_1 \\ k_1/m_2 & -(k_1+k_2)/m_2 & d_1/m_2 & -(d_1+d_2)/m_2 \end{bmatrix} x + \begin{bmatrix} 0 \\ 0 \\ 1/m_1 \\ 0 \end{bmatrix} u$$

If $m_1 = m_2 = 1$, $d_1 = d_2 = 1$ and $k_2 = \frac{1}{4}$, compute the controllability matrix. Reduce this to triangular form using gaussian elimination, and hence determine under what conditions on k_1 the system is controllable.

3.14 The matrix

$$A_n = \begin{bmatrix} 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 \\ -a_n & -a_{n-1} & -a_{n-2} & -a_{n-3} & \dots & -a_1 \end{bmatrix}$$

is said to be in *companion form* because it has characteristic polynomial

$$\det(\lambda I - A_n) = \lambda^n + a_1\lambda^{n-1} + a_2\lambda^{n-2} + \cdots + a_{n-1}\lambda + a_n$$

which can be read off from the last row of A_n.

For example, when $n = 3$

$$A_3 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -a_3 & -a_2 & -a_1 \end{bmatrix}$$

Verify by expanding the determinant that

$$\det(\lambda I - A_3) = \lambda^3 + a_1\lambda^2 + a_2\lambda + a_3$$

Consider the control system model

$$\frac{dx}{dt} = A_n x + du$$

where d is an $n \times 1$ column vector with all entries zero except $d_n = 1$. If linear feedback

$$u = -(f_n x_1 + f_{n-1}x_2 + \cdots + f_2 x_{n-1} + f_1 x_n)$$

is applied obtain the closed loop matrix and verify that it is also in companion form. Hence deduce that the closed loop characteristic polynomial is

$$\lambda^n + (a_1 + f_1)\lambda^{n-1} + (a_2 + f_2)\lambda^{n-2} + \cdots + (a_{n-1} + f_{n-1})\lambda + (a_n + f_n)$$

3.15 For the discrete system

$$x(k+1) = Ax(k), \quad k = 0, 1, 2, \dots$$

with output

$$y(k) = Cx(k)$$

show that

$$Vx(k) = \begin{bmatrix} y(k) \\ y(k+1) \\ \vdots \\ y(k+n-1) \end{bmatrix} \tag{3.100}$$

where V is the observability matrix defined in (3.97).

If

$$A = \begin{bmatrix} 1 & 0 & -1 \\ 1 & 2 & 1 \\ 2 & 2 & 3 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

show that the system is observable. Hence if

$$y(k) = \begin{bmatrix} 2^k \\ 1 - 3(2)^k + 3^k \end{bmatrix}$$

determine $x(k)$ by selecting three independent equations in (3.100) and solving for $x_1(k)$, $x_2(k)$ and $x_3(k)$.

3.16 The *transpose* A^T of a matrix A is obtained by interchanging the rows and columns. For example, if

$$A = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix}$$

then writing the first row as the first column and the second row as the second column gives

$$A^T = \begin{bmatrix} a_1 & a_3 \\ a_2 & a_4 \end{bmatrix}$$

The same holds for vectors:

$$b = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}, \quad b^T = [b_1, b_2, b_3]$$

$$c = [c_1, c_2, c_3], \quad c^T = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}$$

Notice that applying the transpose operation twice brings you back to where you started from, that is

$$(A^T)^T = A, \quad (b^T)^T = b, \quad (c^T)^T = c$$

Consider two systems

$$x(k+1) = Ax(k) + bu(k) \qquad \text{(I)}$$

$$y(k) = cx(k)$$

and

$$x(k+1) = A^T x(k) + c^T u(k) \qquad \text{(II)}$$

$$y(k) = b^T x(k)$$

Denote by U_{I}, V_{I} and U_{II}, V_{II} their respective controllability and observability matrices. Suppose for simplicity that $n = 3$, and show that

$$U_{\text{II}} = V_{\text{I}}^{\text{T}}, \quad V_{\text{II}} = U_{\text{I}}^{\text{T}}$$

You will need the results

$$(Ab)^{\text{T}} = b^{\text{T}}A^{\text{T}}, \quad (A^2)^{\text{T}} = (A^{\text{T}})^2$$

Since controllability of (II) is equivalent to observability of (I), and vice versa, the systems (I) and (II) are called 'dual'.

FURTHER READING

AUSLANDER, D.M., TAKAHASHI, Y. and RABINS, M.J. 1974. *Introducing Systems and Control.* McGraw-Hill, Tokyo.

BARNETT, S. 1990. *Matrices: Methods and Applications.* Oxford University Press, Oxford.

BARNETT, S. and CAMERON, R.G. 1985. *Introduction to Mathematical Control Theory*, 2nd Edition. Oxford University Press, Oxford.

BURGHES, D.N. and DOWNS, A.M. 1975. *Modern Introduction to Classical Mechanics and Control.* Ellis Horwood, Chichester.

BURGHES, D.N. and GRAHAM, A. 1980. *Introduction to Control Theory, including Optimal Control.* Ellis Horwood, Chichester.

FRIEDLAND, B. 1987. *Control System Design.* McGraw-Hill, New York.

MAYR, O. 1970. *The Origins of Feedback Control.* MIT Press, Cambridge, MA.

McCLAMROCH, N.H. 1980. *State Models of Dynamic Systems.* Springer-Verlag, New York.

WIBERG, D.M. 1971. *Theory and Problems of State Space and Linear Systems.* Schaum–McGraw-Hill, New York.

4

All the Best

We remarked at the beginning of Chapter 3 that a basic human desire is to control events – to try and make things happen as we would like them to. This is all part of aiming to 'get the best out of life'. This striving for optimal solutions will be explored in this chapter through a variety of situations.

4.1 SEARCHING FOR AN OPTIMUM

Here we are interested in finding the point at which a function $f(x)$ of one independent variable x achieves its *maximum* or *minimum* value.

■ *EXAMPLE 4.1*

A closed cylindrical beer can is made of sheet metal of constant thickness and is to hold 440 ml. The problem is to find the dimensions of the can so that the amount of metal used in its construction is as small as possible, thereby minimizing its cost.

Suppose the radius of the can is x cm and its height is h cm. The volume is

$$\pi x^2 h = 440$$

so that

$$h = \frac{440}{\pi x^2}$$

The area of each circular flat end is πx^2. The circumference of the curved part is $2\pi x$. If this curved surface is opened out flat it becomes a rectangle of width $2\pi x$ and height h, so its area is $2\pi xh$. Hence the total area of sheet metal is

$$f(x) = 2\pi x^2 + 2\pi xh$$

$$= 2\pi x^2 + 2\pi x\left(\frac{440}{\pi x^2}\right)$$

$$= 2\pi x^2 + \frac{880}{x}$$

We want to find the value of x which minimizes $f(x)$.

To fix ideas, consider the problem of finding the *maximum* value of some given function $f(x)$ of an independent variable x. Suppose it is known that within a certain range of values $[a, b]$ of x, that is $a \leqslant x \leqslant b$, the function

(i) has a *single maximum* point at $x = x^*$, and
(ii) is *unimodal* on this interval $[a, b]$.

Condition (ii) means that if $f(x)$ is a continuous function then it has a single 'hump' as illustrated in Figure 4.1. More precisely, if

$$x_1 < x^* < x_2$$

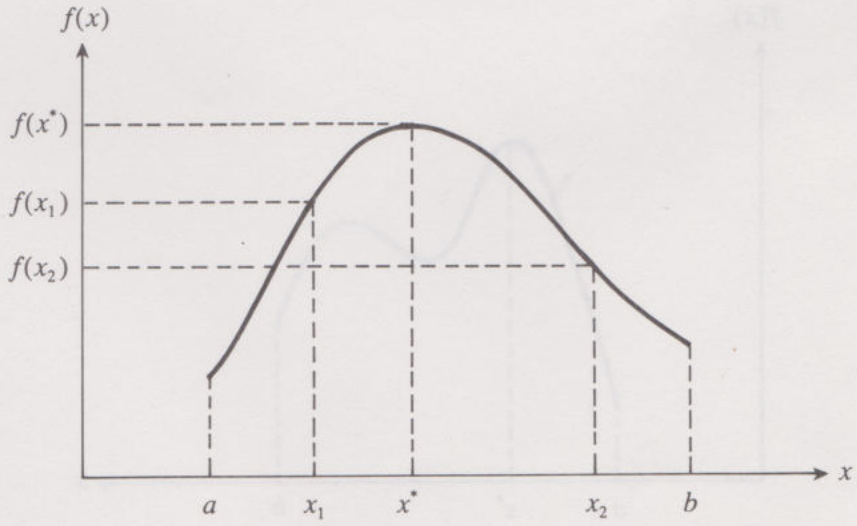


Figure 4.1

then as shown in Figure 4.1 we have

$$f(x_1) < f(x^*), \quad f(x^*) > f(x_2)$$

Notice, however, that a unimodal function need not be continuous – that is, it may contain 'jumps', as illustrated by the example in Figure 4.2. Alternatively a function may not be unimodal as shown by the example in Figure 4.3 where there are two 'humps'. Finding an overall maximum in such cases is more difficult.

In practical applications the function $f(x)$ often represents a profit which is to be maximized. Alternatively, as seen in Example 4.1, the function may represent a cost to be minimized. This causes no difficulties, since minimizing $f(x)$ is the same as maximizing $-f(x)$.

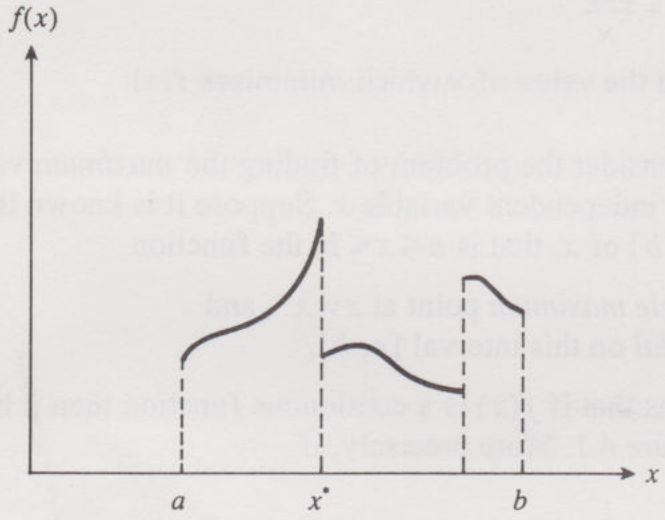


Figure 4.2

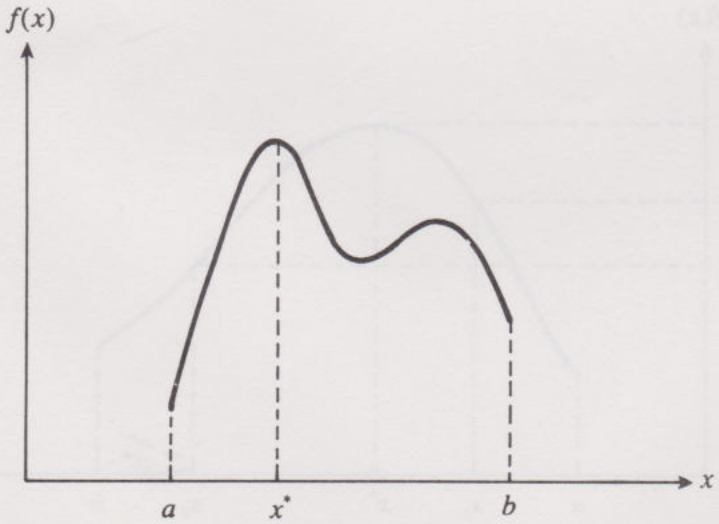


Figure 4.3

EXERCISE 4.1 Find an expression for the rectangular area which can be enclosed by a 500 m length of fencing (let one side of the rectangle be x m). It is required to find the largest possible area.

EXERCISE 4.2 A rectangular sheet of metal 100 cm by 50 cm is to be made into an open rectangular box by cutting out squares of side x cm from each corner as shown in Figure 4.4. and then folding up the sides. Obtain an expression for the volume of the box, which is to be maximized (see Problem 4.1).

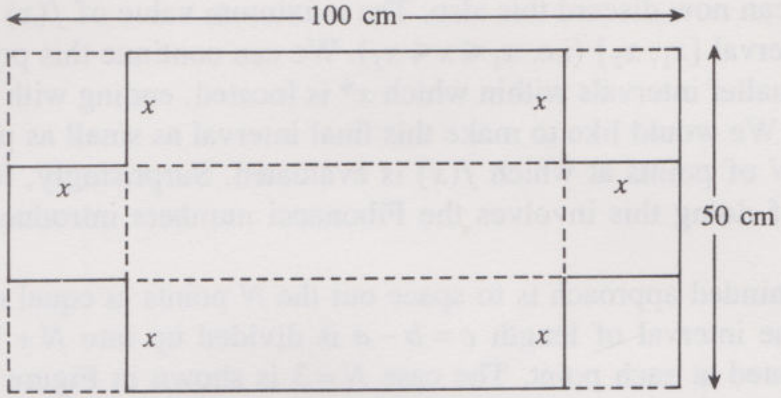


Figure 4.4

If you've studied some calculus you may have encountered a method for finding the maximum (and minimum) values of a function $f(x)$ by determining points at which the derivative df/dx is zero. However, this approach is often inappropriate – it may be that the function has a very complicated formula, or indeed as in Figure 4.2 the derivative may not exist at the maximum point. Alternatively, it may be the case that $f(x)$ is the result of some process, so that we are able only to measure $f(x)$ for certain values of x in the interval $[a, b]$. In all these situations the techniques of calculus are not useful. We therefore need to *search* for the maximum point x^* in some efficient way.

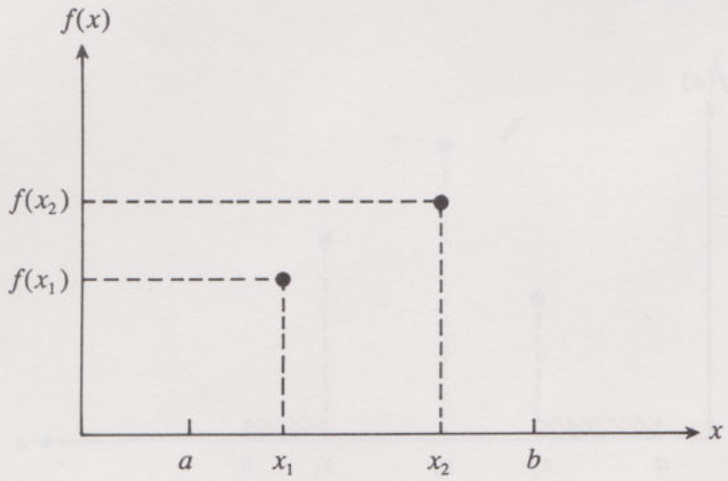


Figure 4.5

Suppose we select two points x_1 and x_2 in the interval $[a, b]$ and find that $f(x_2) > f(x_1)$, as shown in Figure 4.5.

Because $f(x)$ is assumed to be unimodal, it follows that its maximum value cannot occur in the interval $[a, x_1]$, because there cannot be a value of $f(x)$ in $[a, x_1]$ which is bigger than $f(x_2)$. We can therefore disregard this part of the interval when computing further values of $f(x)$, and restrict the next step in our search to the interval $[x_1, b]$. Suppose that for a third point x_3 the situation is as shown in Figure 4.6. By a similar argument it follows that the maximum cannot occur in the interval $[x_2, b]$, so we can now discard this also. The maximum value of $f(x)$ must therefore occur in the interval $[x_1, x_2]$ (i.e. $x_1 \leqslant x \leqslant x_2$). We can continue this procedure to find successively smaller intervals within which x^* is located, ending with a final *interval of uncertainty*. We would like to make this final interval as small as possible using a fixed number N of points at which $f(x)$ is evaluated. Surprisingly, it turns out that the best way of doing this involves the Fibonacci numbers introduced in Example 1.4, Chapter 1.

A simple-minded approach is to space out the N points at equal distances along the interval. The interval of length $c = b - a$ is divided up into $N+1$ equal pieces, and $f(x)$ evaluated at each point. The case $N = 3$ is shown in Figure 4.7. We select the point at which $f(x)$ is largest. In Figure 4.7 this is x_3, so x^* lies in $[x_2, b]$ which has length $2c/4$. In general the final interval of uncertainty has length $2c/(N+1)$.

The Fibonacci process does much better than this. In Chapter 1 we denoted the Fibonacci numbers by

$$f_0 = 1, \quad f_1 = 1, \quad f_2 = 2, \quad f_3 = 3, \quad f_4 = 5, \quad f_5 = 8, \ldots$$

where each number in the sequence is the *sum* of the previous two. When N points are used it turns out that the length of the final interval of uncertainty is $2c/f_{N+1}$. For example, with $N = 6$ this is

$$2c/f_7 = 2c/21$$

compared with $2c/7$ if equidistant spacing is used, and the improvement increases rapidly as N is increased.

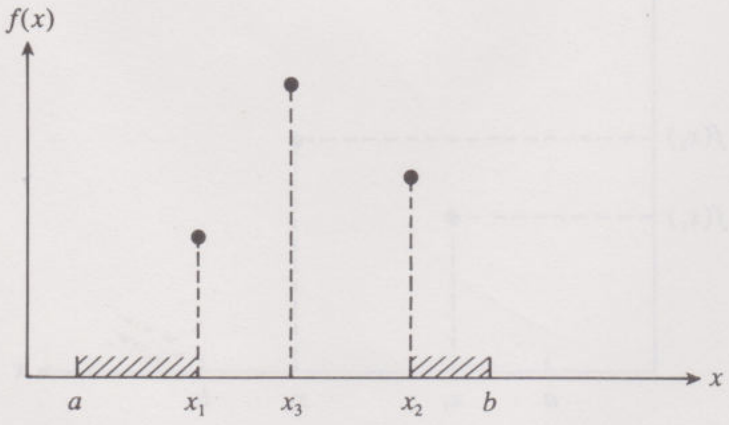


Figure 4.6

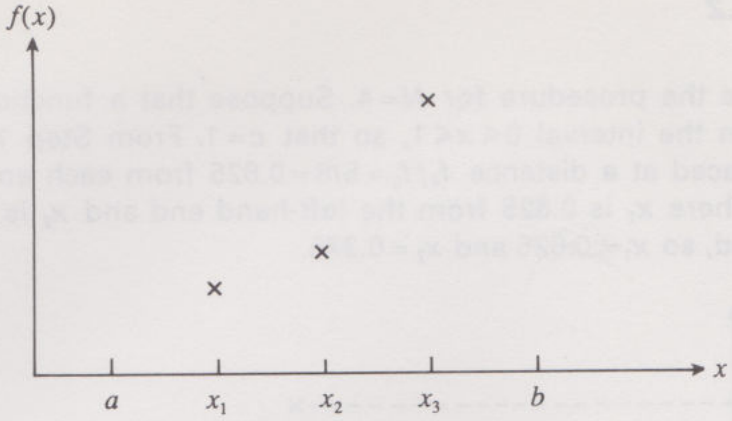


Figure 4.7

The procedure for placing the N points within the interval is as follows:

Fibonacci Search Algorithm

Step 1 Evaluate $f(x)$ at two test points x_1 and x_2 located at distances cf_N/f_{N+1} from each end of the initial interval.

Step 2 According to which of the two values of $f(x)$ is the larger, select the new interval within which x^* must lie.

Step 3 Insert the next point x_3 symmetrically in this new interval with respect to the point already inside it, and evaluate $f(x_3)$.

Step 4 Repeat Steps 2 and 3 until all N points have been inserted. x^* lies within an interval of length $2c/f_{N+1}$.

You might be a little puzzled as to what is meant by 'symmetrically' in Step 3. We say two points P_1,P_2 are symmetrically located in an interval AB if the distance of P_1 from the left-hand end is the same as the distance of P_2 from the right-hand end, as shown in Figure 4.8. It is in fact the symmetric locating of the points in Steps 1 and 3 which makes the method work.

It's worth noting how to modify the search procedure if $f(x)$ is to be *minimized* instead of maximized. It is assumed that $f(x)$ has a single minimum point and is unimodal. Step 1 is unaltered; in the subsequent steps the new interval is selected according to which of the values of $f(x)$ being compared is the *smaller*. For example, if we are looking for a *minimum* of $f(x)$ in Figure 4.5 then the new interval would be $[a, x_2]$.

Figure 4.8

■ EXAMPLE 4.2

Let's illustrate the procedure for $N = 4$. Suppose that a function $f(x)$ is to be maximized on the interval $0 \leq x \leq 1$, so that $c = 1$. From Step 1 the two initial points are placed at a distance $f_4/f_5 = 5/8 = 0.625$ from each end, as shown in Figure 4.9, where x_1 is 0.625 from the left-hand end and x_2 is 0.625 from the right-hand end, so $x_1 = 0.625$ and $x_2 = 0.375$.

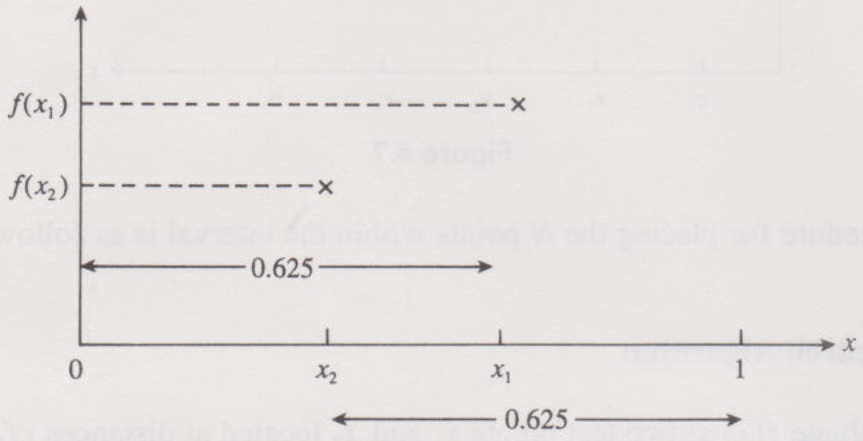


Figure 4.9

These two points are symmetrically located in the interval [0, 1]. Suppose $f(x_1) > f(x_2)$, so in Step 2 we delete the interval $[0, x_2]$ since we are assuming that $f(x)$ is unimodal. In Step 3 the next point x_3 is placed 0.375 from the left-hand end of the new interval $[x_2, 1]$, so that it is symmetrically located relative to the interior point x_1, as shown in Figure 4.10.

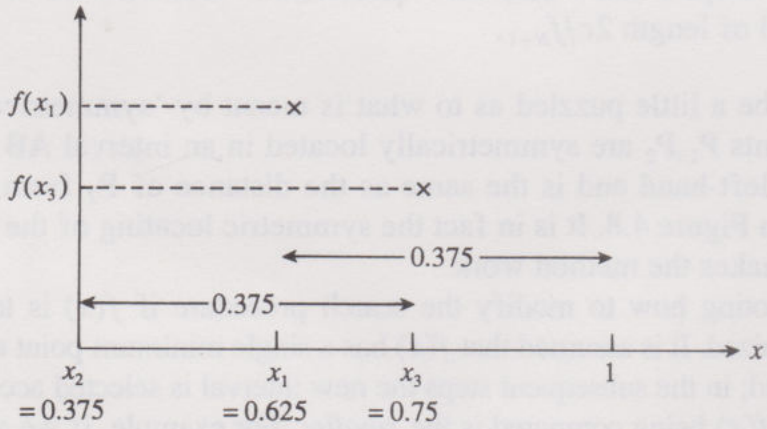


Figure 4.10

If we suppose $f(x_3) < f(x_1)$ then the next new interval is $[x_2, x_3]$. The final point x_4 is located at a distance

$$x_3 - x_1 = 0.75 - 0.625 = 0.125$$

from the left-hand end, as shown in Figure 4.11.

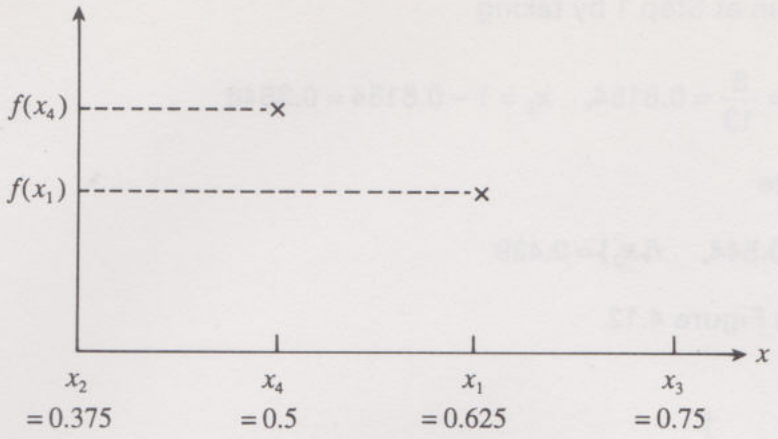


Figure 4.11

If we suppose $f(x_4) > f(x_1)$ then we drop the interval $[x_1, x_3]$. The final interval of uncertainty is therefore $[x_2, x_1]$.

The following table shows the end points of the successive intervals within which x^* lies:

	End points		Length of interval
Initial interval	0	1	1
After first step	0.375	1	0.625
Second step	0.375	0.75	0.375
Third step	0.375	0.625	0.25

You can see that the final interval of uncertainty has length $0.625 - 0.375 = 0.25$, which agrees with the theoretical value of $2c/f_5 = 2/8$. Notice also that, reading downwards in the last column, the length of each interval is the *sum* of the two below, that is

$$1 = 0.625 + 0.375$$
$$0.625 = 0.375 + 0.25$$

This rule is exactly the same as that used to generate Fibonacci numbers, and this is essentially the reason why they arise in this context.

■ *EXAMPLE 4.3*

Given the function

$$f(x) = x(1.5 - x), \quad 0 \leqslant x \leqslant 1$$

we apply Fibonacci search with five evaluations of $f(x)$ (i.e. $N = 5$) to estimate the value of x^* which maximizes $f(x)$. The length of the final interval of uncertainty within which x^* lies will be

$$2c/f_6 = 2/13 = 0.1538$$

We begin at Step 1 by taking

$$x_1 = \frac{f_5}{f_6} = \frac{8}{13} = 0.6154, \quad x_2 = 1 - 0.6154 = 0.3846$$

and compute

$$f(x_1) = 0.544, \quad f(x_2) = 0.429$$

as shown in Figure 4.12.

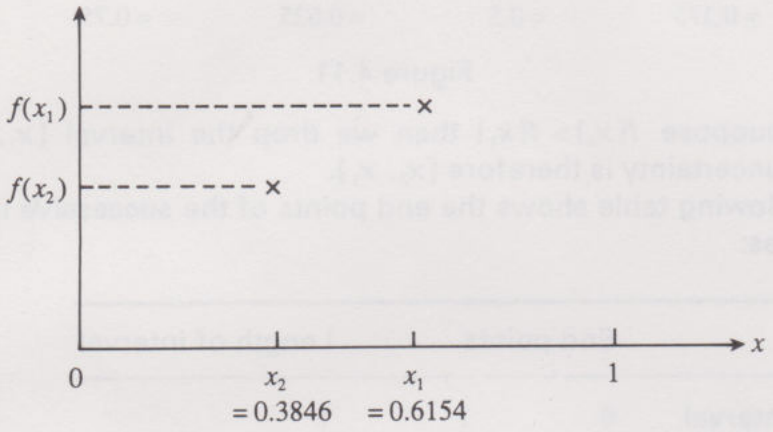


Figure 4.12

In Step 2 we therefore select the new interval $[x_2, 1]$ and in Step 3 we take (see Figure 4.13)

$$x_3 = 1 - 0.2308 = 0.7692$$

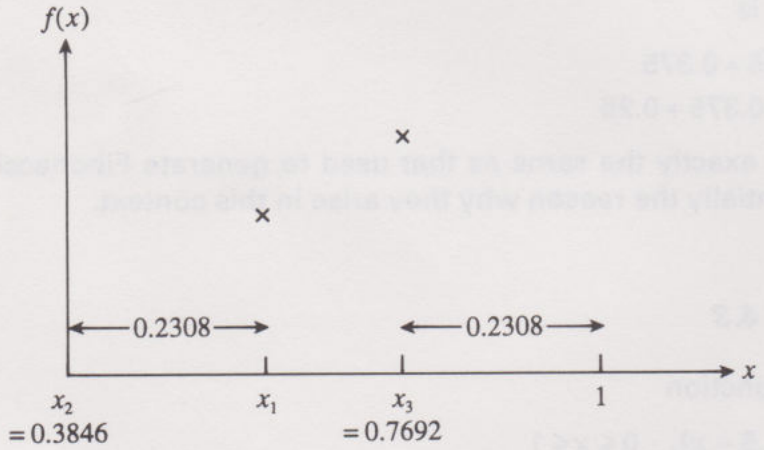


Figure 4.13

which produces $f(x_3) = 0.562 > f(x_1)$. We next select the interval $[x_1, 1]$ and take $x_4 = 0.8462$ (see Figure 4.14) giving $f(x_4) = 0.553 < f(x_3)$.

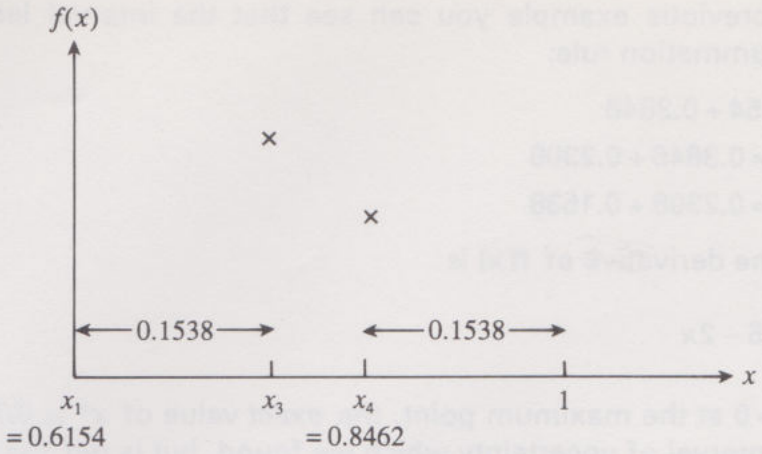


Figure 4.14

The next new interval is $[x_1, x_4]$ and we take $x_5 = 0.6924$ (see Figure 4.15) with $f(x_5) = 0.559 < f(x_3)$.

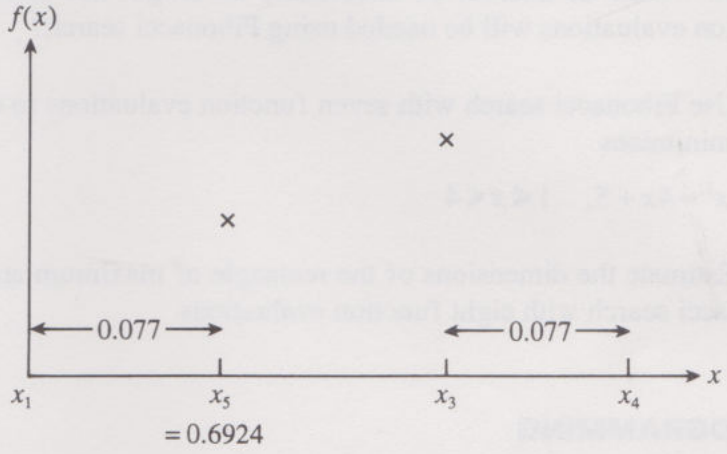


Figure 4.15

The final interval of uncertainty is therefore $[x_5, x_4]$, that is $0.6924 \leqslant x \leqslant 0.8462$, which has length 0.1538, agreeing with the theoretical value. The midpoint of this interval is 0.7693, so we can say that $x^* = 0.7693 \pm 0.0769$. Again it is instructive to list the successive intervals of uncertainty with their lengths:

	End points		Length of interval
Initial interval	0	1	1
$[x_2, 1]$	0.3846	1	0.6154
$[x_1, 1]$	0.6154	1	0.3846
$[x_1, x_4]$	0.6154	0.8462	0.2308
$[x_5, x_4]$	0.6924	0.8462	0.1538

As in the previous example you can see that the interval lengths obey the Fibonacci summation rule:

$1 = 0.6154 + 0.3846$

$0.6154 = 0.3846 + 0.2308$

$0.3846 = 0.2308 + 0.1538$

Since the derivative of $f(x)$ is

$$\frac{df}{dx} = 1.5 - 2x$$

and $df/dx = 0$ at the maximum point, the *exact* value of x^* is 0.75. This does lie within the interval of uncertainty which we found, but is *not* the midpoint.

EXERCISE 4.3 Repeat Example 4.3 using $N = 6$.

EXERCISE 4.4 If the maximum value of a unimodal function $f(x)$ defined on $1 \leqslant x \leqslant 5$ is to be located within an interval of uncertainty of length not more than 0.05, how many function evaluations will be needed using Fibonacci search?

EXERCISE 4.5 Use Fibonacci search with seven function evaluations to estimate the value of x which minimizes

$$f(x) = x^2 - 4x + 5, \quad 1 \leqslant x \leqslant 4$$

EXERCISE 4.6 Estimate the dimensions of the rectangle of maximum area in Exercise 4.1 using Fibonacci search with eight function evaluations.

4.2 LINEAR PROGRAMMING

■ *EXAMPLE 4.4*

The manager of a small convenience store decides to stock two brands of ice cream, A and B, but believes that likely sales will only justify ordering a total of at most 25 cartons. The space available in the freezer for ice cream is at most 36 cubic units. Because of different packaging, a carton of brand A occupies 1 cubic unit, but brand B takes up 2 cubic units per carton. The profit of brand A is 60 pence per carton and on B 90 pence per carton. How much of each brand should be stocked so as to make the maximum profit?

Let x_1, x_2 denote the numbers of cartons of brands A and B respectively which are stocked. The *constraint* on the total number of cartons is

$$x_1 + x_2 \leqslant 25 \tag{4.1}$$

The volume occupied by the cartons is $x_1 + 2x_2$ so the *constraint* on available space is

$$x_1 + 2x_2 \leqslant 36 \tag{4.2}$$

Notice that because of the way they are defined both x_1 and x_2 are *non-negative*, that is

$$x_1 \geqslant 0, \quad x_2 \geqslant 0 \tag{4.3}$$

We wish to find the values of x_1 and x_2 which maximize the profit, which is (in pence)

$$z = 60x_1 + 90x_2 \tag{4.4}$$

This is a typical example of what is called a *linear programming* (*LP*) problem. This name was invented in 1951 when 'programming' was a fashionable new scientific term, much like 'fractals' or 'chaos' today. Computer programming was a very new discipline, and terms like 'mathematical programming', 'quadratic programming' and 'dynamic programming' were coined to emphasize their novelty, and also their need for computers to perform multitudinous computations. The description *linear* refers to the fact that the expressions on the left-hand sides of the inequalities in (4.1) and (4.2) are linear combinations of the variables x_1 and x_2, and so is the profit function z in (4.4). We pointed out in Section 3.3 in Chapter 3 that a *linear combination* of quantities $x_1, x_2, x_3, \ldots$ simply means the expression

$$a_1x_1 + a_2x_2 + a_3x_3 + \cdots$$

where $a_1, a_2, a_3, \ldots$ are constants, not all zero. In this example, however, we can also show how 'linear' relates to the geometrical meaning of *straight line*. Let's look at the inequality (4.1): when *equality* holds we have

$$x_1 + x_2 = 25 \tag{4.5}$$

which is the straight line AB in Figure 4.16(a), remembering that we are restricted to the quadrant $x_1 \geqslant 0$, $x_2 \geqslant 0$.

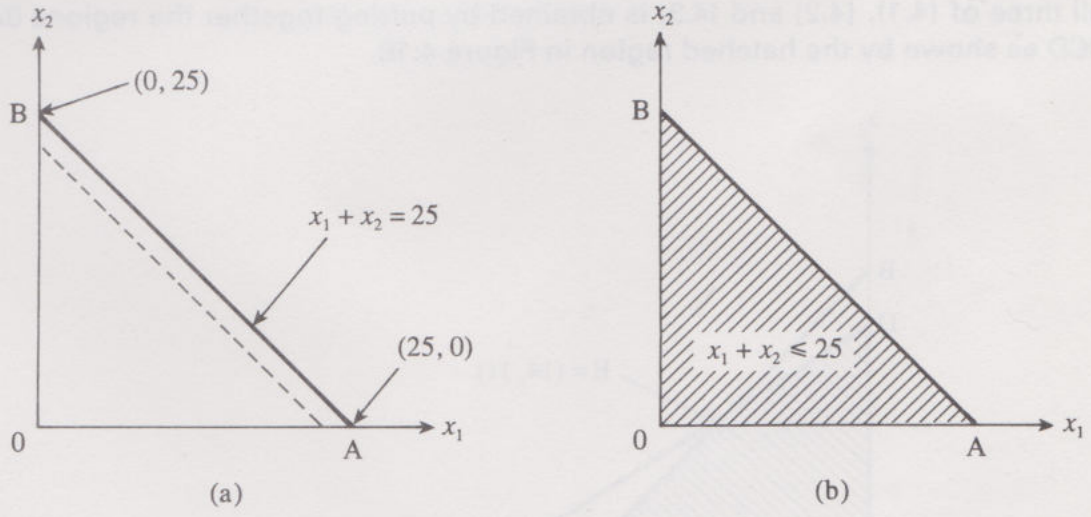


Figure 4.16

Consider next the equation

$$x_1 + x_2 = 24 \ (< 25)$$

This equation corresponds to the dashed line in Figure 4.16(a). Similarly, all equations of the form

$$x_1 + x_2 = k$$

with $k < 25$ are represented by parallel lines *inside* the triangle 0AB. Remembering that we cannot have x_1 or x_2 negative, we have therefore found that the inequality (4.1) corresponds to the hatched region 0AB in Figure 4.16(b). Using an exactly similar argument, the inequality (4.2) together with (4.3) corresponds to the hatched region 0CD shown in Figure 4.17.

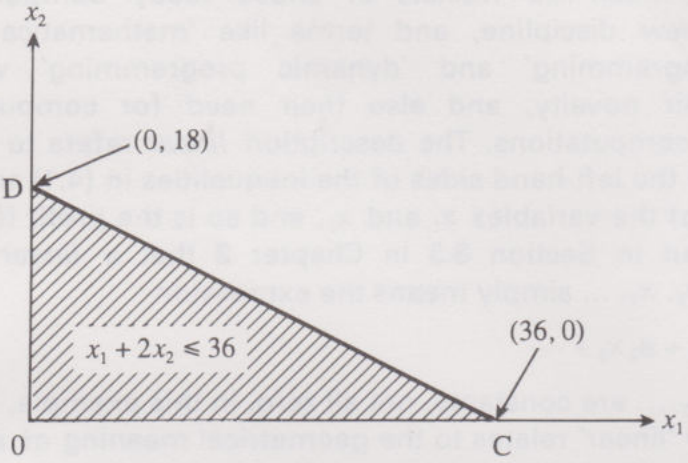


Figure 4.17

By the way, it's easy to find the coordinates of the vertices A, B, C, D. Simply set $x_2 = 0$ in (4.5) to get $A = (25, 0)$, and similarly $x_1 = 0$ gives $B = (0, 25)$; C and D are found from $x_1 + 2x_2 = 36$ in the same way. The overall region which satisfies all three of (4.1), (4.2) and (4.3) is obtained by putting together the regions 0AB, 0CD as shown by the hatched region in Figure 4.18.

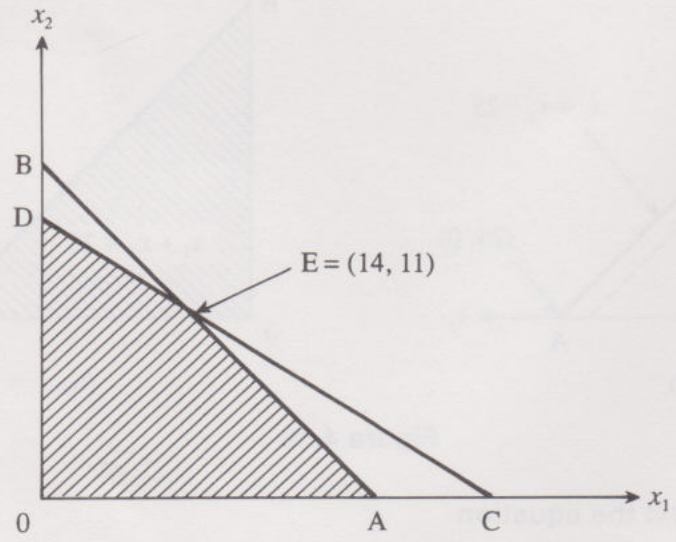


Figure 4.18

The point E is where the two lines AB, CD intersect, and its coordinates are found by solving the simultaneous equations

$$x_1 + x_2 = 25, \quad x_1 + 2x_2 = 36$$

to give E = (14, 11). The region 0AED is called the *feasible region* since it contains all the points which satisfy the constraints (4.1), (4.2) and (4.3). Any such point is called a *feasible solution* to the LP problem.

The profit function z in (4.4) can also be represented in terms of straight line graphs. For example, if $z = 180$ we have the straight line

$$60x_1 + 90x_2 = 180$$

which is shown in Figure 4.19 joining the points (0, 2) and (3, 0). As the value of z increases we obtain a series of parallel lines which move away from the origin: the case $z = 360$ is also shown in Figure 4.19.

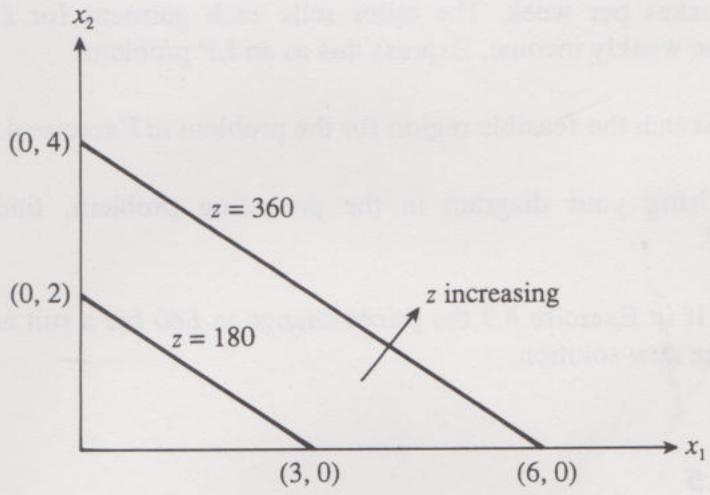


Figure 4.19

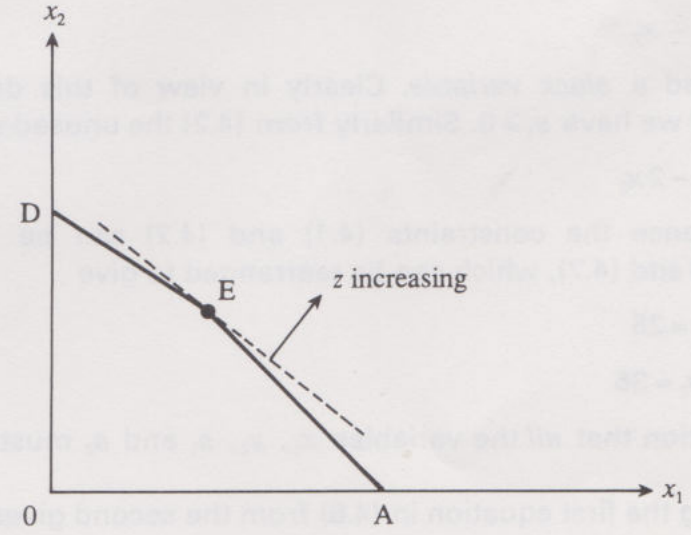


Figure 4.20

Our LP problem can now be stated in geometrical terms: what is the furthest from the origin that the 'z-line' can be whilst still remaining in the feasible region? The answer is shown in Figure 4.20: the maximum value of z is attained when the dashed line passes through the vertex E, which has coordinates $x_1 = 14$, $x_2 = 11$. That is, stocking 14 cartons of brand A and 11 of brand B produces the maximum possible profit

$$z = 60 \times 14 + 90 \times 11 = 1830 \text{ pence}$$

You may be thinking that it's fortunate that the solution came out in integers – how could the store stock a fractional number of cartons? This difficulty will be discussed later, in the next section.

EXERCISE 4.7 A tailor has 80 m^2 of polyester material and 120 m^2 of wool material available each week. A suit requires 1 m^2 of polyester and 3 m^2 of wool, whereas a dress requires 2 m^2 of each. Let x_1 and x_2 be the numbers of suits and dresses which the tailor makes per week. The tailor sells each garment for £50, and wishes to maximize the weekly income. Express this as an LP problem.

EXERCISE 4.8 Sketch the feasible region for the problem in Exercise 4.7.

EXERCISE 4.9 Using your diagram in the preceding problem, find the solution for Exercise 4.7.

EXERCISE 4.10 If in Exercise 4.7 the prices change to £80 for a suit and £40 for a dress, determine the new solution.

■ *EXAMPLE 4.5*

Let's now tackle the problem in Example 4.4 using an *algebraic* approach. The total number of cartons stocked is $x_1 + x_2$, so the number of unfulfilled sales is

$$s_1 = 25 - x_1 - x_2 \qquad (4.6)$$

and s_1 is called a *slack variable*. Clearly in view of this definition and the constraint (4.1) we have $s_1 \geqslant 0$. Similarly from (4.2) the unused storage space is

$$s_2 = 36 - x_1 - 2x_2 \qquad (4.7)$$

with $s_2 \geqslant 0$. Hence the constraints (4.1) and (4.2) can be replaced by the equations (4.6) and (4.7), which can be rearranged to give

$$x_1 + x_2 + s_1 = 25 \qquad (4.8)$$

$$x_1 + 2x_2 + s_2 = 36$$

with the condition that *all* the variables x_1, x_2, s_1 and s_2 must be non-negative ($\geqslant 0$).

Subtracting the first equation in (4.8) from the second gives

$$x_2 = 11 + s_1 - s_2 \qquad (4.9)$$

From the first equation in (4.8) we then get

$$\begin{aligned} x_1 &= 25 - x_2 - s_1 \\ &= 14 - 2s_1 + s_2, \quad \text{using (4.9)} \end{aligned} \tag{4.10}$$

Substituting the expressions (4.9) and (4.10) into the profit function in (4.4) gives

$$\begin{aligned} z &= 60x_1 + 90x_2 \\ &= 60(14 - 2s_1 + s_2) + 90(11 + s_1 - s_2) \\ &= 1830 - 30s_1 - 30s_2 \end{aligned} \tag{4.11}$$

Since s_1 and s_2 cannot be negative, the maximum value of z in (4.11) must occur when $s_1 = 0$, $s_2 = 0$, in which case $z = 1830$. From (4.9) and (4.10) we then have $x_1 = 14$, $x_2 = 11$. This agrees with the result found graphically (see Figure 4.20).

Sometimes an inequality constraint arises the other way round, for example we might have

$$x_1 + 3x_2 \geqslant 9$$

In this case we define the corresponding slack variable by

$$s = x_1 + 3x_2 - 9$$

so that $s \geqslant 0$, and the inequality is replaced by the equation

$$x_1 + 3x_2 - s = 9$$

The crucial thing is to keep *all* the variables non-negative.

EXERCISE 4.11 Solve algebraically the LP problem:

$$\begin{aligned} &\text{Maximize} && z = 2x_1 + 3x_2 \\ &\text{Subject to:} && 2x_1 + 4x_2 \leqslant 13 \\ & && 3x_1 + 2x_2 \leqslant 11 \\ & && x_1 \geqslant 0, \quad x_2 \geqslant 0 \end{aligned}$$

EXERCISE 4.12 Solve algebraically the LP problem:

$$\begin{aligned} &\text{Minimize} && z = 3x_1 + 2x_2 \\ &\text{Subject to:} && x_1 + 3x_2 \geqslant 9 \\ & && 2x_1 + x_2 \geqslant 14 \\ & && x_1 \geqslant 0, \quad x_2 \geqslant 0 \end{aligned}$$

When there are more than two variables the graphical technique cannot be used. However, a powerful approach called the *simplex method* can be applied (using appropriate software) to solve LP problems with hundreds of variables and constraints. The basic ideas of the method are illustrated in the following example.

■ EXAMPLE 4.6

Consider the LP problem:

$$\text{Maximize} \quad z = 5x_1 - x_2 \tag{4.12}$$

$$\begin{aligned} \text{Subject to:} \quad & x_1 + x_2 \leqslant 5 \\ & 2x_1 + 3x_2 \leqslant 11 \\ & -x_1 + 2x_2 \leqslant 2 \\ & x_1 \geqslant 0, \quad x_2 \geqslant 0 \end{aligned} \tag{4.13}$$

We first write the inequalities (4.13) as equations by introducing slack variables $s_1 \geqslant 0$, $s_2 \geqslant 0$, $s_3 \geqslant 0$, so that

$$\begin{aligned} x_1 + x_2 + s_1 &= 5 \\ 2x_1 + 3x_2 + s_2 &= 11 \\ -x_1 + 2x_2 + s_3 &= 2 \end{aligned} \tag{4.14}$$

You must remember that throughout none of the variables can be negative. Since there are three equations in (4.14) we select three of the variables and solve for them in terms of the other two variables. A *basic* solution is one where three variables are positive and the other two are zero.

(i) Choose x_1, x_2, s_1, which gives

$$\begin{aligned} x_1 &= \frac{16}{7} - \frac{2s_2}{7} + \frac{3s_3}{7} \\ x_2 &= \frac{15}{7} - \frac{s_2}{7} - \frac{2s_3}{7} \\ s_1 &= \frac{4}{7} + \frac{3s_2}{7} - \frac{s_3}{7} \end{aligned} \tag{4.15}$$

Substituting into (4.12) produces

$$z = \frac{65}{7} - \frac{9s_2}{7} + \frac{17s_3}{7} \tag{4.16}$$

A *basic* solution to the problem is obtained by taking $s_2 = 0$, $s_3 = 0$ in (4.15), which produces $x_1 = 16/7$, $x_2 = 15/7$ and $s_1 = 4/7$. In this case $z = 65/7$. However, we can see from (4.16) that this solution is not optimal, since by making s_3 positive we can increase the value of z. Keeping $s_2 = 0$, (4.15) gives us

$$x_1 = \frac{1}{7}(16 + 3s_3), \quad x_2 = \frac{1}{7}(15 - 2s_3), \quad s_1 = \frac{1}{7}(4 - s_3) \tag{4.17}$$

Since x_1, x_2 and s_1 in (4.17) are not allowed to be negative, you can see that the *largest* value of s_3 which we can take is $s_3 = 4$. In this case $s_1 = 0$, which means that in the trio $\{x_1, x_2, s_1\}$ of positive variables s_1 is now replaced by s_3:

(ii) The selected variables are now x_1, x_2 and s_3. Re-solve the original equations (4.14) to obtain

$$\begin{aligned} x_1 &= 4 - 3s_1 + s_2 \\ x_2 &= 1 + 2s_1 - s_2 \\ s_3 &= 4 - 7s_1 + 3s_2 \end{aligned} \qquad (4.18)$$

Substituting these expressions into the profit function (4.12) gives

$$z = 19 - 17s_1 + 6s_2 \qquad (4.19)$$

Hence a second basic solution is obtained by taking $s_1 = 0$, $s_2 = 0$, which gives $x_1 = 4$, $x_2 = 1$, $s_3 = 4$ and $z = 19$. Notice that this value of z is larger than before. Again we see that z in (4.19) can be made yet larger, this time by increasing the value of s_2. Looking at the expressions (4.18) you can see that with $s_1 = 0$, the largest possible value of s_2 is $s_2 = 1$, in which case $x_2 = 0$. We therefore replace x_2 by s_2 in the trio of positive variables, giving:

(iii) The selected variables are now x_1, s_2, s_3. The solution of (4.14) is

$$\begin{aligned} x_1 &= 5 - x_2 - s_1 \\ s_2 &= 1 - x_2 + 2s_1 \\ s_3 &= 7 - 3x_2 - s_1 \end{aligned}$$

and

$$z = 25 - 6x_2 - 5s_1$$

You can see from this last expression that z will achieve its maximum value, equal to 25, when $x_2 = 0$, $s_1 = 0$. The values of the other variables are then $x_1 = 5$, $s_2 = 1$, $s_3 = 7$ and we have found the optimal solution.

It's instructive to look at a graphical interpretation of the algebraic work we've just done in steps (i), (ii) and (iii). The feasible region described by the inequalities (4.13) is the hatched area 0ABCD in Figure 4.21. This diagram is built up in the same way as Figure 4.18 in Example 4.4. The profit function z in (4.12) is represented by the dashed line in Figure 4.21, and is obtained by the same arguments used in Figure 4.19. For example, when $z = 10$ the line represented by (4.12) passes through F = (2, 0) and G = (0, −10). When $z = 20$ the line is further away from the origin in the direction of the arrow in Figure 4.21.

For step (i) the basic solution is $s_2 = 0$, $s_3 = 0$ which is the point B in Figure 4.21; at the next step we move to $s_1 = 0$, $s_2 = 0$ which is the point C; finally, we move to $s_1 = 0$, $x_2 = 0$ which is the point D.

As shown in Figure 4.22, each basic solution corresponds to a *corner* of the feasible region. The key idea of the simplex method is that we move from one corner to another in such a way that the profit function z is increased, as shown by the dashed parallel lines in Figure 4.22. This continues until it is not possible to

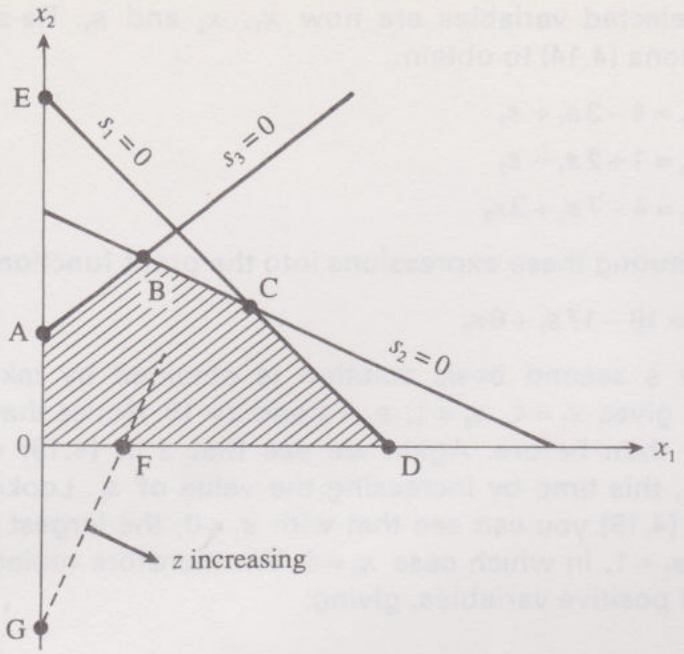


Figure 4.21

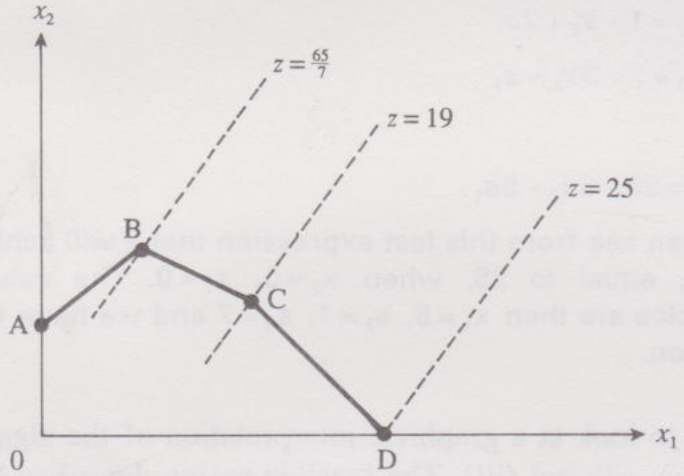


Figure 4.22

increase z any further – in this example, we cannot go beyond the point D without going outside the feasible region.

EXERCISE 4.13 Solve the following LP problems graphically, and by using the simplex procedure.

Subject to: $x_1 + x_2 \leqslant 3$
$x_1 - 2x_2 \leqslant 1$
$-2x_1 + x_2 \leqslant 2$
$x_1 \geqslant 0, \quad x_2 \geqslant 0$

(a) maximize $z = x_1 - x_2$,
(b) minimize $z = x_1 - x_2$.

In practice the arithmetical operations involved in carrying out the simplex method can be made purely mechanical by expressing them in a tabular form. Details of this are rather tedious and can be found in textbooks listed at the end of the chapter. To solve a practical problem you would use standard available software.

4.3 TRANSPORTATION MODELS

A special form of LP problem involves the transportation of goods from a set of depots to a set of destinations, the aim being to do this as cheaply as possible.

■ *EXAMPLE 4.7*

Suppose there are four warehouses, which are to supply three supermarkets with certain goods. Each supermarket requires 5 units, but the amounts available at the warehouses are respectively 1, 6, 2, 6 units: notice that the total supply does equal the total demand (15 units). The cost of transporting one item from a warehouse to a supermarket is shown in the following table:

		Warehouses			
		(1)	(2)	(3)	(4)
Supermarkets	(1)	5	4	3	2
	(2)	10	8	4	7
	(3)	9	9	8	4

(4.20)

The number c_{ij} in row i, column j in (4.20) is the cost of transporting 1 unit to supermarket (i) from warehouse (j). For example, the unit transportation cost to supermarket (2) from warehouse (4) is $c_{24} = 7$. Let's denote by x_{ij} the number of units which come to supermarket i from warehouse j, as shown in (4.21).

		Availabilities			
		1	6	2	6
Requirements	5	x_{11}	x_{12}	x_{13}	x_{14}
	5	x_{21}	x_{22}	x_{23}	x_{24}
	5	x_{31}	x_{32}	x_{33}	x_{34}

(4.21)

For the first supermarket, the incoming total is required to be 5 units, so adding up the first row gives

$$x_{11} + x_{12} + x_{13} + x_{14} = 5 \qquad (4.22)$$

An exactly similar operation applies for each of the other two supermarkets, so adding the elements in the other two rows gives

$$\begin{aligned} x_{21} + x_{22} + x_{23} + x_{24} &= 5 \\ x_{31} + x_{32} + x_{33} + x_{34} &= 5 \end{aligned} \tag{4.23}$$

Considering the goods which are despatched, the amounts going out from the first depot are given in the first column, so adding these gives

$$x_{11} + x_{21} + x_{31} = 1 \tag{4.24}$$

Similarly for the other three depots we must have

$$\begin{aligned} x_{12} + x_{22} + x_{32} &= 6 \\ x_{13} + x_{23} + x_{33} &= 2 \\ x_{14} + x_{24} + x_{34} &= 6 \end{aligned} \tag{4.25}$$

Notice that the constraints this time are equations instead of inequalities, but obviously we must still have all $x_{ij} \geqslant 0$ – we can't transport a negative number of items! To find the total cost z which is to be minimized, we multiply each unit cost given in (4.20) by the corresponding amount transported in (4.21), to end up with

$$\begin{aligned} z = {} & 5x_{11} + 4x_{12} + 3x_{13} + 2x_{14} \\ & + 10x_{21} + 8x_{22} + 4x_{23} + 7x_{24} \\ & + 9x_{31} + 9x_{32} + 8x_{33} + 4x_{34} \end{aligned} \tag{4.26}$$

You can see that the transportation problem is a special kind of LP problem: we wish to minimize the *linear* cost function (4.26) subject to the *linear* constraints (4.22), (4.23), (4.24) and (4.25). In fact, one of these seven equations is redundant, because of the condition that the total of units available is equal to the total number required. This means that the expression obtained by adding together the three equations (4.22) and (4.23) is identical to that obtained by adding together the four equations (4.24) and (4.25).

However, because of its special form we don't use the simplex method to solve transportation problems, but instead develop a special procedure. First, to find a *feasible* solution (i.e. one which satisfies all the constraints) we apply the *northwest corner* method, so called because we start in the 'northwest corner' in (4.21) with the cell containing x_{11}. We wish to fill in the cells in the array (4.27).

		Availabilities			
		1	6	2	6
Requirements	5				
	5				
	5				

(4.27)

This is to be done in such a way that the row sums and column sums add up to the correct amounts. It's clear that the most we can put into the (1, 1) cell is 1, because of the 1 at the top of the first column; the supply from the first warehouse is therefore exhausted, so no more entries can go into the first column. However, supermarket (1) still requires 4 units, so we move along the first row to the next cell (i.e. the northwest corner of the remaining array), and see that we can supply all 4 units from warehouse (2). At this stage we therefore have the array in (4.28):

	1	6	2	6
5	1 ①	4 ②	0	0
5	0			
5	0			

(4.28)

The slashed-through numbers indicate that these are finished with; the numbers within the circles indicate the sequence in which we have filled the cells. We now put the maximum possible amount into the northwest corner of the array which is left; clearly this is 2 in the (2, 2) cell. We need a total of 5 units in the second row, but can only put 2 into the (2, 3) cell because only 2 units remain available for the second column. To complete this second row, we therefore need 1 unit in the (2, 4) cell, giving the array in (4.29):

	1	6	2	6
5	1 ①	4 ②	0	0
5	0	2 ③	2 ④	1 ⑤
5	0	0	0	

(4.29)

The only way to complete the solution is to put 5 units into the (3, 4) cell, giving the feasible solution shown in (4.30).

	1	6	2	6
5	1	4	0	0
5	0	2	2	1
5	0	0	0	5

(4.30)

The cost of this solution is obtained from (4.26) as

$$z = 5\times 1 + 4\times 4 + 8\times 2 + 4\times 2 + 7\times 1 + 4\times 5$$
$$= 72$$

A glance at the table of costs in (4.20) reveals that the two cheapest routes (namely, supermarket (1) from warehouses (3) and (4)) are not used in the solution

(4.30). It therefore seems sensible to try a scheme which starts with the cheapest route, assigns as much as possible to this, then uses the next cheapest route available, and so on. The result of this procedure is shown in the array (4.31), where again the numbers within circles indicate the sequence in which the cells are filled. However, using (4.26) you can easily check that the cost of this solution is $z = 82$, which is dearer than before! The 'common sense' approach to finding a solution by concentrating on the cheapest routes therefore won't work in general.

	1	6	2	6
5	0	0	0	5 ①
5	0	3 ④	2 ③	0
5	1 ⑥	3 ⑤	0	1 ②

(4.31)

Clearly, a more systematic method is needed to find the best solution. Begin with the first solution displayed in (4.30), and define *shadow costs* u_1, u_2, u_3, v_1, v_2, v_3, v_4. These variables are chosen so that the cost c_{ij} in (4.20) for each cell (i, j) in (4.30) which contains a non-zero entry satisfies

$$c_{ij} = v_i + u_j \tag{4.32}$$

You can think of the v as 'dispatch' costs, and the u as 'reception' costs, so (4.32) says that each cost for an occupied cell is the sum of a dispatch cost and a reception cost. For the solution in (4.30) the situation is as follows:

Reception at supermarkets

u_1	5	4		
u_2		8	4	7
u_3				4
	v_1	v_2	v_3	v_4

Dispatch from warehouses

By (4.32) the shadow costs must satisfy the equations

$$u_1 + v_1 = 5, \quad u_1 + v_2 = 4$$
$$u_2 + v_2 = 8, \quad u_2 + v_3 = 4, \quad u_2 + v_4 = 7$$
$$u_3 + v_4 = 4$$

Since there are *seven* unknowns but only *six* equations the solution is not unique. We can arbitrarily set any one of the variables, say v_1, equal to zero and solve for the rest. The solution in this case for the shadow costs is shown in (4.33). Notice that

u_1	5	5	4		
u_2	9		8	4	7
u_3	6				4
		0	−1	−5	−2
		v_1	v_2	v_3	v_4

(4.33)

some can be negative. Next, form the array whose entry for each *empty* cell (i, j) in (4.33) is $u_i + v_j - c_{ij}$, where the c_{ij} are the costs in (4.20). We obtain the table (4.34), where for example the (1, 3) entry is

$$u_1 + v_3 - c_{13} = 5 - 5 - 3 = -3$$

		−3	1
−1			
−3	−4	−7	

(4.34)

The total cost will be reduced if we transport items via the route which corresponds to the *positive* entry in (4.34), namely from warehouse (4) to supermarket (1). This will replace *one* of the routes used in the original solution (4.30). We transfer as many as possible into this new cell (1, 4) without making any other entry negative. Inspection of (4.30) shows that we can put at most 1 unit into cell (1, 4), otherwise the entry in (2, 4) becomes negative. Our improved solution is now

	1	6	2	6
5	1	3	0	1
5	0	3	2	0
5	0	0	0	5

(4.35)

and new shadow costs are shown in (4.36):

u_1	5	5	4		2
u_2	9		8	4	
u_3	7				4
		0	−1	−5	−3
		v_1	v_2	v_3	v_4

(4.36)

The table $u_i + v_j - c_{ij}$ for empty cells in (4.36) is shown in (4.37). The entries in

		–3	
–1			–1
–2	–3	–6	

(4.37)

are all negative, showing that the solution in (4.35) is optimal. The cost in (4.26) is now $z = 71$, which is the smallest possible.

An explanation of why the method works is as follows. At each step we calculate the entries in the table $u_i + v_j - c_{ij}$ for cells which are empty in the current solution. If there are any positive entries we select the largest, and put as much as possible into the corresponding cell. Suppose a cell which is *not* currently utilized as a transportation route is cell (2, 1). If we enter 1 unit into this cell, then in order to balance we have the situation in (4.38), where the other three cells shown *are* occupied in the current solution – that is, cells $(2, k)$, $(j, 1)$ and (j, k) contain non-zero entries.

$$\begin{array}{lcccc} \text{row } 2 & u_2 & \boxed{1} & \cdots & -1 \\ & \vdots & \vdots & \cdots & \vdots \\ \text{row } j & u_j & -1 & \cdots & 1 \\ & & v_1 & & v_k \\ & & \text{column 1} & & \text{column k} \end{array} \qquad (4.38)$$

The change in the cost due to the changes in the solution shown in (4.38) is

$$c_{21} - c_{2k} - c_{j1} + c_{jk} = c_{21} - (u_2 + v_k) - (u_j + v_1) + (u_j + v_k) \qquad (4.39)$$

$$= c_{21} - (u_2 + v_1) \qquad (4.40)$$

where (4.39) follows from the definition (4.32) of the shadow costs for occupied cells. You can see from (4.40) that there will be an improvement (i.e. the cost is reduced) if $c_{21} < u_2 + v_1$. Hence we use cell (2, 1) if $u_2 + v_1 - c_{21} > 0$. The procedure is repeated, selecting one new cell to be occupied at each iteration, until there are no positive entries in the table $u_i + v_j - c_{ij}$.

One point you should be aware of: for both LP and transportation problems, although the optimal value of the cost (or profit) function is unique, there may be more than one set of values of the variables which produces this – hence we look for *an* optimal solution, as it may not be unique.

EXERCISE 4.14 Use the northwest corner method to find a feasible solution of the transportation problem with the following requirements and availabilities:

		Availabilities		
		15	25	5
Requirements	5			
	15			
	10			
	15			

EXERCISE 4.15 A transportation problem with the cost table

7	3
9	4
10	5

has the feasible solution

		Availabilities 3	5
	4	0	4
Requirements	2	1	1
	2	2	0

Use the shadow costs method to obtain an optimal solution, and show that the minimum cost is 42.

EXERCISE 4.16 Consider a transportation problem with availabilities, requirements and costs as shown in the following table.

		Availabilities 4	6	10
	3	3	2	4
Requirements	5	5	7	6
	12	3	4	5

Obtain a feasible solution using the northwest corner method. Use the shadow costs method to obtain an optimal solution, and show that the minimum cost is 85.

You can see that the solution method developed for the transportation problem guarantees that the variables always have integer values, assuming the availabilities and requirements are integers. This is fortunate, since it would not make sense to transport fractional quantities of items. However, for general LP problems there is no such guarantee that the simplex method will produce a solution in integers, even if this is required for the solution to be valid. You were correct if you thought that the way the solution to Example 4.4 came out in integers was highly fortuitous. Another example where the answer luckily comes out in integers is provided by Exercises 4.7, 4.8 and 4.9, in which a tailor wishes to find the most profitable scheme for making suits and dresses. Problem 4.4 involving an allocation of buses provides yet a further example where the solution fortunately works out in integers. Finding optimal solutions in integer form for general LP problems is a subject called *integer programming*, which is beyond the scope of this book. Instead we end this section with some examples of special LP problems where integer solutions can be found relatively easily.

EXAMPLE 4.8

The 'assignment problem' is a special case of the transportation problem. The basic idea is that there are n 'candidates' to be allocated to n 'jobs', and this is to be done so that the overall 'cost' is optimized.

(a) Four people P_1, P_2, P_3, P_4 are to be assigned to each carry out one of four tasks T_1, T_2, T_3, T_4. The 'fitness' of person P_i for job T_j is evaluated by some assessment procedure to have a value c_{ij}, with $c_{ij}=0$ if the person cannot do this task. Let $x_{ij}=1$ if P_i is assigned to T_j, and $x_{ij}=0$ otherwise. For example, the number of jobs done by P_1 is $x_{11}+x_{12}+x_{13}+x_{14}$, and since each person does only one job we can write

$$x_{11}+x_{12}+x_{13}+x_{14}=1$$

The same argument holds for P_2, P_3 and P_4, so we have

$$x_{i1}+x_{i2}+x_{i3}+x_{i4}=1, \quad i=1,2,3,4 \tag{4.41}$$

Similarly, the number of people doing job T_j is

$$x_{1j}+x_{2j}+x_{3j}+x_{4j}=1, \quad j=1,2,3,4 \tag{4.42}$$

It is required to maximize the overall effectiveness of the assignment, which is

$$\sum_i \sum_j c_{ij}x_{ij} = c_{11}x_{11}+c_{12}x_{12}+\cdots+c_{43}x_{43}+c_{44}x_{44} \tag{4.43}$$

subject to the constraints (4.41) and (4.42).

(b) A university wants to advertise four of its degree programmes in four different periodicals. The cost varies according to the size of the advertisement and the publication, and is c_{ij} for advertisement i in periodical j. Set $x_{ij}=1$ if course i is advertised in periodical j, and $x_{ij}=0$ otherwise. The overall cost is again (4.43), and in this case is to be minimized. Assuming that the university can afford to place just one advertisement for each course, the constraints are again (4.41) and (4.42).

(c) Four firms of building contractors are asked to give quotations for the costs of carrying out four jobs. However, each firm only has the resources to undertake one of the contracts. Let c_{ij} be the quotation from contractor i for job j and let $x_{ij}=1$ if contractor i does job j, and $x_{ij}=0$ otherwise. The problem is again to minimize the overall cost (4.43) subject to the constraints (4.41) and (4.42).

EXAMPLE 4.9

Five towns are linked by motorways as shown in Figure 4.23.

A supermarket chain has a store in each of the towns. The chain decides to supply each store from a depot which is either within the same town, or in a town with which there is a direct motorway link. They want to have as few

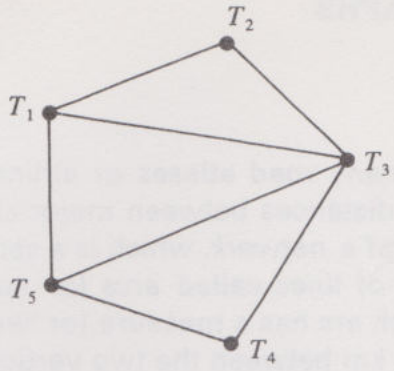


Figure 4.23

depots as possible. Let $x_i = 1$ if town T_i contains a depot, and $x_i = 0$ if it does not. From Figure 4.23 you can see that town T_1 has direct links with towns T_2, T_3 and T_5. Hence the number of depots to which T_1 has access is

$$x_1 + x_2 + x_3 + x_5$$

which includes the possibility that there is a depot in T_1 itself. The requirement that T_1 must be served by at least one depot means that

$$x_1 + x_2 + x_3 + x_5 \geqslant 1$$

Similarly for the other towns the constraints are

T_2: $x_1 + x_2 + x_3 \geqslant 1$

T_3: $x_1 + x_2 + x_3 + x_4 + x_5 \geqslant 1$

T_4: $x_3 + x_4 + x_5 \geqslant 1$

T_5: $x_1 + x_3 + x_4 + x_5 \geqslant 1$

We have to minimize the total number of depots

$$x_1 + x_2 + x_3 + x_4 + x_5$$

In fact, the answer is that there need be only two depots – one solution is to have depots in T_2 and T_4. Can you spot any other solutions?

EXERCISE 4.17 An athletics competition consists of four track events: 100 m, 400 m, 800 m, 1500 m. The rules are that each runner from a club can enter only one event, and the team with the smallest total time for the four events is the winner. The trial times in seconds of four athletes in a club are as follows:

		Event			
		100 m	400 m	800 m	1500 m
Athlete	A	13	64	146	385
	B	12	62	147	365
	C	14	63	150	360
	D	14	61	145	370

Express in assignment form the problem of selecting who runs in each event so as to produce the best result for the club.

4.4 NETWORKS AND GRAPHS

■ *EXAMPLE 4.10*

A common feature of many road atlases or airline route maps is a diagrammatic representation of distances between major cities, as illustrated in Figure 4.24. This is an example of a *network*, which is a set of points called *vertices* (or *nodes*) and a collection of lines called *arcs* (or *edges*) joining some or all of these points in pairs. Each arc has a measure (or 'weight') attached to it – in this example, the distance in km between the two vertices. Incidentally, the drawing need not be to scale. If there are no numbers associated with the arcs then the network is given the technical name *graph* (e.g. see Figure 4.23). This is a completely different usage of the term from what you are familiar with as the graph of a function.

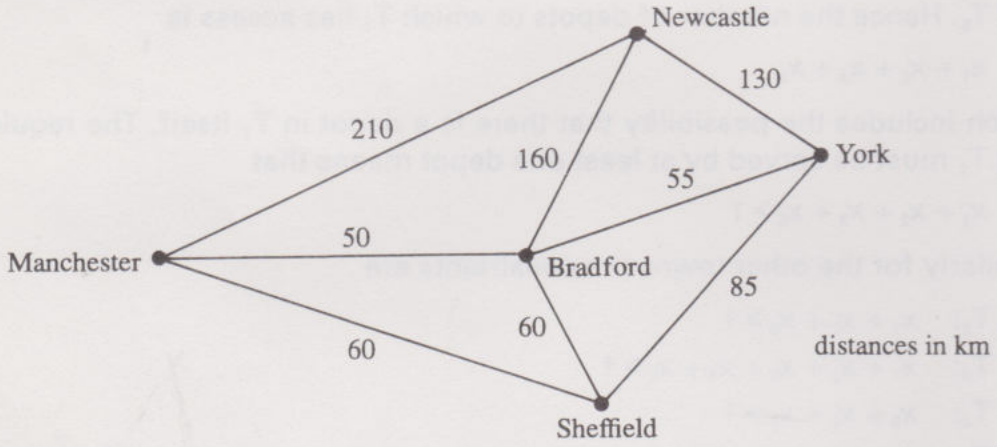


Figure 4.24

■ *EXAMPLE 4.11*

Graphs are often useful in giving a pictorial representation of a situation. Let's look at a simple example of an assignment problem like those in Section 4.3, but without any optimization involved. Suppose there are five teachers (T_1 to T_5) to be timetabled to take five classes in Mathematics, Science, Economics, History and French. The capabilities of the teachers are

Teacher	Can teach
T_1	Mathematics and Science
T_2	Mathematics and French
T_3	Mathematics and Economics
T_4	History and Economics
T_5	French and History.

The graphical representation of this information is shown in Figure 4.25. An arc connects a teacher to a subject if it is one they can teach. The graph in Figure 4.25

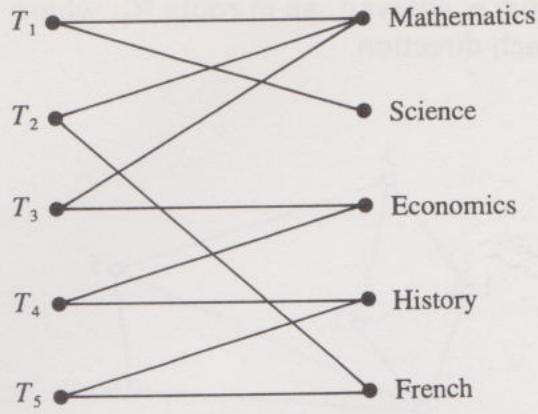


Figure 4.25

is called *bipartite*, since there are two collections (or 'subsets') of vertices (the teachers and the subjects) which have no arcs connecting vertices in the same subset (there are no arcs connecting teachers, and no arcs connecting subjects). An assignment problem is to find a set of arcs which connects just one teacher to each subject.

EXERCISE 4.18 Find two satisfactory assignments in Example 4.11.

In a road network like Figure 4.24 it's usually the case that the arcs can be travelled in either direction. However, you can imagine a city-centre plan where many of the streets are one way to traffic. If the arcs have directional arrows on them then the graph is called *directed*, but in this section we'll only consider *undirected* graphs where there are no arrows on any of the arcs. A convenient notation is to use (i, j) to denote an arc connecting vertex i to vertex j. For example, consider the graph in Figure 4.23. The arc connecting vertices T_1 and T_2 is $(1, 2)$, and the entire graph can be described by the set of arcs

$\{(1, 2), (1, 3), (1, 5), (2, 3), (3, 4), (3, 5), (4, 5)\}$

■ *EXAMPLE 4.12*

(a) Looking again at Figure 4.24, you can see that there may be several different routes joining two cities. For example, to go from Manchester to Sheffield we could use any of the following routes:

R_1: Manchester–Bradford–Newcastle–York–Sheffield
R_2: Manchester–Bradford–Newcastle–York–Bradford–Sheffield
R_3: Manchester–Bradford–York–Bradford–Sheffield

The route R_1 is an example of a *path*, where no vertex is visited more than once. The routes R_2 and R_3 are not paths, since in each case Bradford is passed through twice. They are examples of what is called a *walk*, which is simply any set of arcs connecting one vertex to another. Notice that

'retracing of steps' is allowed, as in route R_3, where the arc Bradford–York is traversed in each direction.

(b)

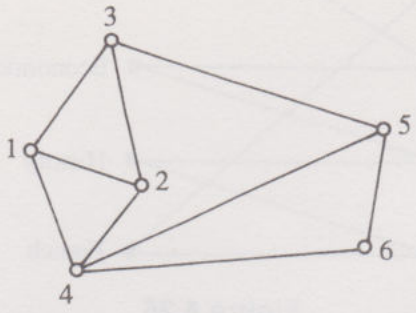


Figure 4.26

In Figure 4.26 the vertices are labelled simply with integers.

(i) The set of arcs

{(1, 2), (2, 4), (4, 5), (5, 6)}

is a path from vertex 1 to vertex 6.

(ii) The set of arcs

{(1, 3), (3, 5), (5, 4), (4, 1)}

is a *cycle*, which is a path which starts and finishes at the same vertex (here vertex 1).

(iii) The set of arcs

{(1, 2), (2, 4), (4, 2), (2, 3)}

is a walk from vertex 1 to vertex 3.

EXERCISE 4.19

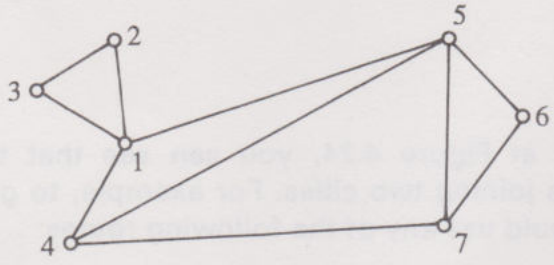


Figure 4.27

What is the nature of the following graphs in Figure 4.27?

(a) {(1, 3), (3, 2), (2, 1)}
(b) {(1, 2), (2, 3), (3, 1), (1, 4)}
(c) {(5, 6), (6, 7), (7, 4), (4, 1), (1, 2)}

For a road network like that in Figure 4.24 we're often interested in finding the shortest path between two given cities. It's obvious by looking at Figure 4.24 that none of the routes R_1, R_2 or R_3 is any good in this respect – the shortest distance between Manchester and Sheffield is clearly 60 km. What we need for a general network is a systematic way of finding a shortest path between two vertices s (= start) and f (= finish). This is provided by *Dijkstra's method.*

We assign a 'label' $L(i)$ to every vertex i in the network, equal to the distance to that vertex from the starting vertex s along the shortest path found so far. The label may be permanent (P), in which case the problem is solved for that vertex; otherwise the label is temporary (T), in which case there is uncertainty as to whether the path to this vertex from s is the shortest possible.

We begin with a set of vertices all holding temporary labels denoted by '∞', which represents a very large number, much larger than any of the numbers assigned to the arcs of the network.

At each step we reduce by *one* the number of vertices with temporary labels. This is done by finding paths to these vertices using the shortest path to vertices with permanent labels, followed by an arc from such a vertex. The vertex with the *smallest* temporary label is then made permanent. The procedure is successively repeated, making one new label permanent each time, until the final vertex f receives a permanent label.

Dijkstra's Algorithm

Step 1 Set $L(s) = 0$, $L(i) = \infty$ for $i \neq s$.
Let p = last vertex to be given a permanent label.
Set $p = s$.

Step 2 For each vertex i with a temporary label, compute its new label using the formula

$$\underset{\text{new label}}{L(i)} = \min[\underbrace{L(i), L(p)}_{\text{old labels}} + d(p, i)] \tag{4.44}$$

In (4.44), $d(p, i)$ is the length of the arc (p, i), and the labels within the square brackets are the values at the *previous* iteration.
Find the vertex k with the smallest new temporary label.
Set $p = k$ and make $L(p)$ permanent.

Step 3 If vertex f has a temporary label, repeat Step 2. Otherwise, the solution has been obtained.

Actually using the algorithm is not as complicated as it seems in the formal description!

■ EXAMPLE 4.13

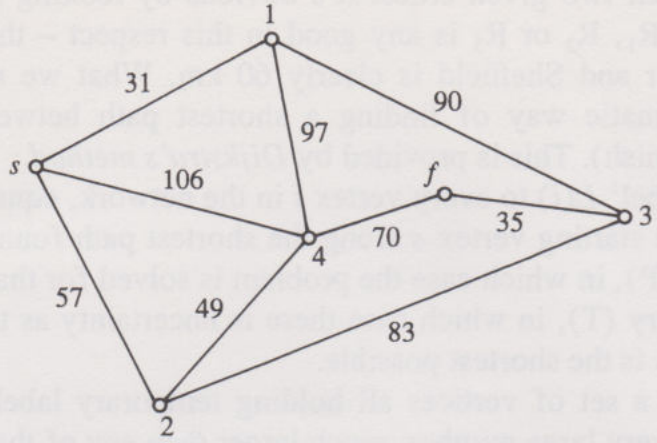


Figure 4.28

Let's find a shortest path between the vertices s and f for the network in Figure 4.28. The algorithm proceeds as follows:

Step 1 This assigns the initial labels

Vertex	s	1	2	3	4	f
Label $L(i)$	0	∞	∞	∞	∞	∞
Status	P	T	T	T	T	T

and we have $p = s$.

Step 2 We now redefine the labels $L(i)$ using (4.44). With $i = 1$ we get

$$L(1) = \min[L(1), L(s) + d(s, 1)]$$
$$= \min[\infty, 0 + 31] = 31$$

Notice that we get the values of $L(1)$ and $L(s)$ inside the square brackets from the table in Step 1. The value of $d(s, 1)$ comes from Figure 4.28. We repeat this process for $i = 2, 3, 4, f$ as follows:

$$L(2) = \min[L(2), L(s) + d(s, 2)]$$
$$= \min[\infty, 0 + 57] = 57$$
$$L(3) = \min[L(3), L(s) + d(s, 3)]$$
$$= \min[\infty, 0 + \infty] = \infty$$

Notice that we denote $d(s, 3)$ by ∞, since there is no arc joining vertices s and 3.

$$L(4) = \min[L(4), L(s) + d(s, 4)]$$
$$= \min[\infty, 0 + 106] = 106$$
$$L(f) = \min[L(f), L(s) + d(s, f)]$$
$$= \min[\infty, 0 + \infty] = \infty$$

The smallest of these new temporary labels is $L(1)=31$, so $k=1$. We therefore set $p=1$ and make $L(1)$ permanent. The updated table of labels is

Vertex	s	1	2	3	4	f
Label $L(i)$	0	31	57	∞	106	∞
Status	P	P	T	T	T	T

Since vertex f has a temporary label, we repeat Step 2 of the algorithm for those vertices having a temporary label. Using (4.44) now gives

$$L(2) = \min[L(2), L(1) + d(1,2)]$$
$$= \min[57, 31 + \infty] = 57$$

$$L(3) = \min[L(3), L(1) + d(1,3)]$$
$$= \min[\infty, 31 + 90] = 121$$

$$L(4) = \min[L(4), L(1) + d(1,4)]$$
$$= \min[106, 31 + 97] = 106$$

$$L(f) = \min[L(f), L(1) + d(1,f)]$$
$$= \min[\infty, 31 + \infty] = \infty$$

The smallest of these values is $L(2)=57$ so we set $p=2$, make $L(2)$ permanent, and the new table is

Vertex	s	1	2	3	4	f
Label $L(i)$	0	31	57	121	106	∞
Status	P	P	P	T	T	T

We again repeat Step 2 of the algorithm.
The new labels are

$$L(3) = \min[L(3), L(2) + d(2,3)]$$
$$= \min[121, 57 + 83] = 121$$

$$L(4) = \min[L(4), L(2) + d(2,4)]$$
$$= \min[106, 57 + 49] = 106$$

$$L(f) = \min[L(f), L(2) + d(2,f)]$$
$$= \min[\infty, 57 + \infty] = \infty$$

We see that $p = 4$, so $L(4)$ is made permanent and the new table is

Vertex	s	1	2	3	4	f
Label $L(i)$	0	31	57	121	106	∞
Status	P	P	P	T	P	T

EXERCISE 4.20 Repeat Step 2 of the algorithm two more times to produce the final table:

Vertex	s	1	2	3	4	f
Label $L(i)$	0	31	57	121	106	156
Status	P	P	P	P	P	P

We have therefore found that the shortest path from s to f in Figure 4.28 has length 156, since by definition this is the value of $L(f)$ when $L(f)$ has become permanent.

EXERCISE 4.21 Use Dijkstra's algorithm to find the length of the shortest path from vertex s to vertex f for the network shown in Figure 4.29.

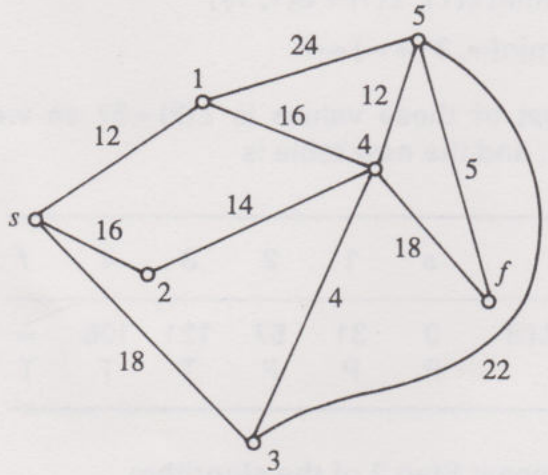


Figure 4.29

The algorithm as described so far gives only the length of the shortest path. To find the shortest path itself we add another step to the algorithm, enabling us to find the vertices on the optimal path.

Step 4 For *each* permanently labelled vertex j other than the starting vertex s, define a vertex $r(j)$ as follows:

$$r(j) = i, \quad \text{where } L(j) = L(i) + d(i, j), \quad i \neq j \tag{4.45}$$

If this is not unique it means there is more than one shortest path. A shortest path, in *reverse order*, is

$$[r(f), f] = (a_1, f), [r(a_1), r(f)] = (a_2, a_1)$$
$$[r(a_2), r(a_1)] = (a_3, a_2), [r(a_3), r(a_2)] = (a_4, a_3), \ldots \quad (4.46)$$

as shown in Figure 4.30.

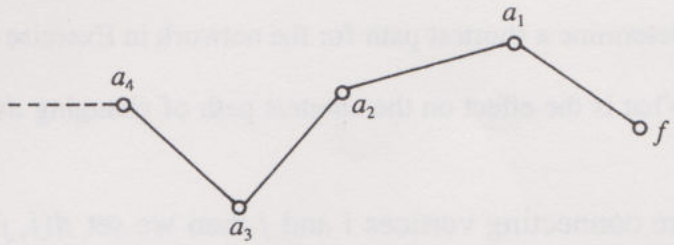


Figure 4.30

■ ***EXAMPLE 4.13 (continued)***

Use the final table of labels, set out in Exercise 4.20. In (4.45) take $j = 1$ and refer to Figure 4.28 to obtain

$r(1) = s$, since $L(1) = 31 = L(s) + d(s, 1)$

With $j = 2$ in (4.45) we get

$r(2) = s$, since $L(2) = 57 = L(s) + d(s, 2)$

and similarly

$j = 3$: $r(3) = 1$, since $L(3) = 121 = L(1) + d(1, 3)$

$j = 4$: $r(4) = 2$, since $L(4) = 106 = L(2) + d(2, 4)$

$j = f$: $r(f) = 3$, since $L(f) = 156 = L(3) + d(3, f)$

Using (4.46), the *last* arc of the shortest path is

$[r(f), f] = (3, f)$

Equation (4.46) shows that the previous arc is

$[r(3), r(f)] = (1, 3)$

Repeat the process to get the next previous arc:

$[r(1), r(3)] = (s, 1)$

and we have now reached the starting vertex. The shortest path is therefore as shown in Figure 4.31.

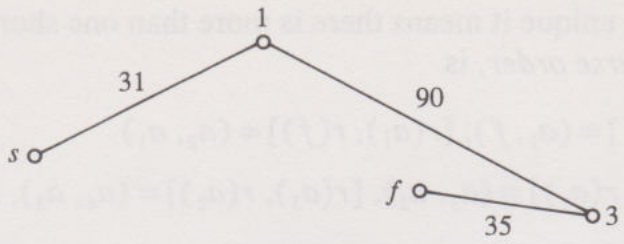


Figure 4.31

EXERCISE 4.22 Determine a shortest path for the network in Exercise 4.21.

EXERCISE 4.23 What is the effect on the shortest path of changing $d(5, f)$ from 5 to 6 in Figure 4.29?

If there is no arc connecting vertices i and j then we set $d(i, j) = \infty$ in Dijkstra's algorithm. In the same way we can apply the method to a directed network by setting $d(i, j) = \infty$ if there is no directed arc from i to j.

We now turn to an interesting type of graph called *trees*. The name arises from the idea of a 'family tree' which shows the relationships between generations – a family tree for a bee population is shown in Figure 1.5 in Chapter 1.

■ *EXAMPLE 4.14*

The United Kingdom postcode consists of two letters signifying the postal town, for example LS stands for Leeds. There are 120 postcode areas. This is

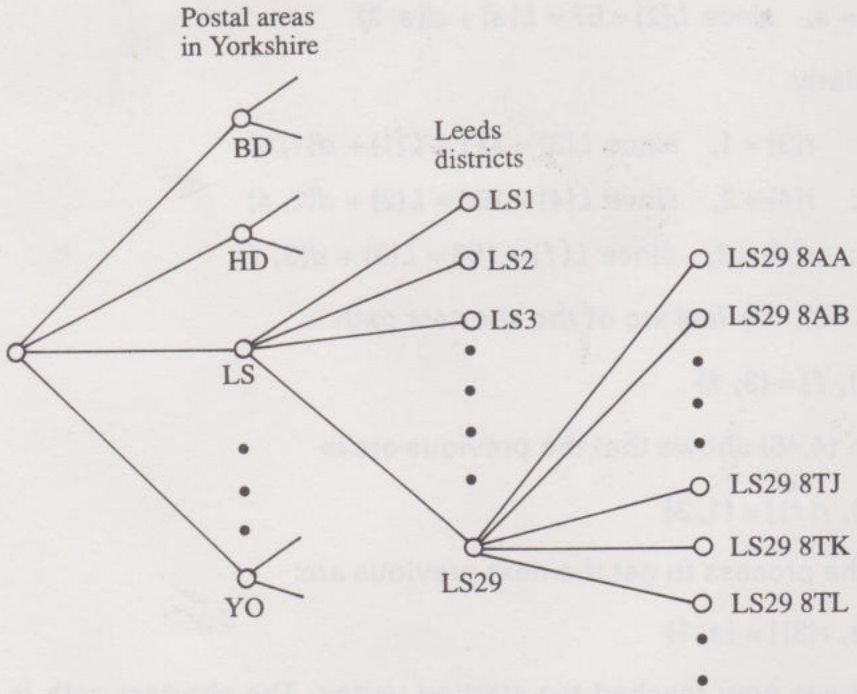


Figure 4.32

followed by one or two digits denoting the postcode district, for example LS1 is part of central Leeds, LS22 is Wetherby and LS29 is Ilkley. Finally, one or two digits denote the sector within the district, and two more alphabetical characters pinpoint the address to within 15 letterboxes on average. The diagram in Figure 4.32 is a *tree* which shows how an item sent to a destination with postcode LS29 8TL can be sorted in stages.

A graph is called *connected* if there is a path from each vertex to every other vertex. The precise definition of a *tree* is a connected graph which does not contain any cycles (defined in Example 4.12(b)). Trees have the following properties:

(i) If there are n vertices then there are $n-1$ arcs.
(ii) Every pair of vertices is linked by exactly one path.
(iii) The removal of any arc produces a *disconnected* graph (i.e. one which is not connected).

Properties (ii) and (iii) are direct consequences of the definition of a tree. Property (i) requires a proof by induction, and can be found in textbooks on graphs. It is interesting, however, that a converse result holds which enables us to recognize trees: if a graph has no cycles, n vertices and $n-1$ arcs then it is a tree (see Problem 4.14).

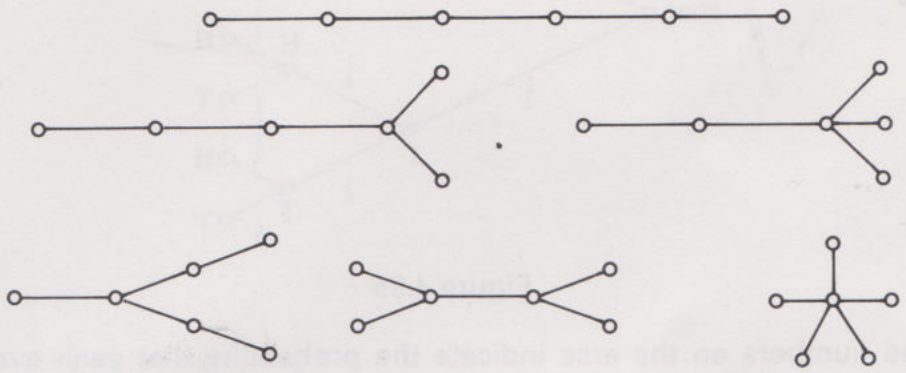


Figure 4.33

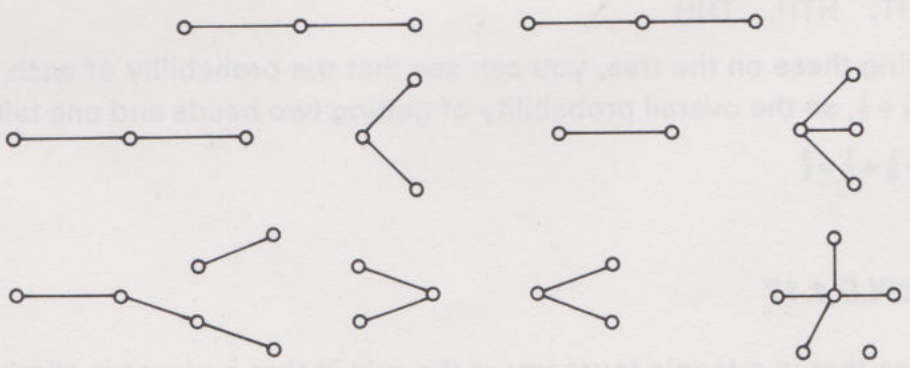


Figure 4.34

■ EXAMPLE 4.15

All possible trees having six vertices are shown in Figure 4.33. You can see that they each have five arcs. Also, there is a path connecting every pair of vertices; if one arc is removed then we are left with a disconnected graph, some examples of this being given in Figure 4.34 for each of the trees in Figure 4.33.

Trees can be used to count all the possible ways in which a sequence of events can occur.

■ EXAMPLE 4.16

An unbiased coin is tossed three times. The various possible outcomes are shown by the tree in Figure 4.35. We read from left to right, and H indicates the coin comes up heads, T that it shows tails.

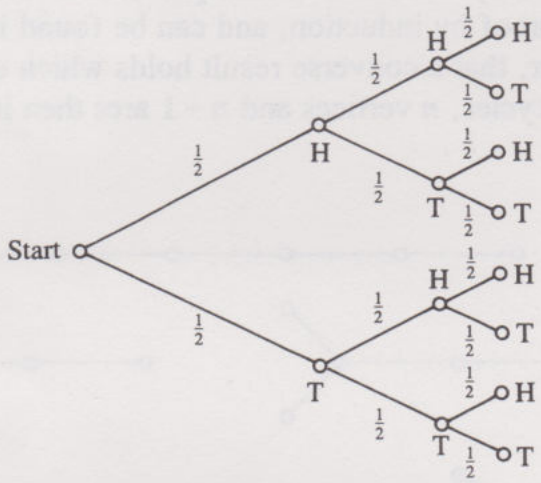


Figure 4.35

The numbers on the arcs indicate the probability that each event occurs, and these are all $\frac{1}{2}$, since the coin is unbiased. For example, two heads and one tail can be obtained in three ways:

HHT, HTH, THH

Following these on the tree, you can see that the probability of each of them is $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}$, so the overall probability of getting two heads and one tail is

$$\frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8}$$

■ EXAMPLE 4.17

Suppose that in a tennis tournament the rule is that a player is eliminated after a single defeat. Let's look at the problem of how to find the smallest number of

matches which need to be played so that there is an overall winner. To represent the tournament by a tree, start with the last match, the winner of which will be the overall champion. The players in this last match must have been the winners of two previous matches as represented in Figure 4.36.

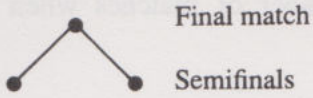


Figure 4.36

Similarly each of the four semifinalists must have been the winner of four previous matches, as shown in Figure 4.37. At *each stage* of the tournament the number of black vertices is the number of matches played, and the number of players involved is therefore twice this. For example, at the quarter-final stage in Figure 4.37 there are four matches and eight players. Thus if there were exactly eight players, the situation shown in Figure 4.37 would be sufficient, and there would be a total of $4 + 2 + 1 = 7$ matches.

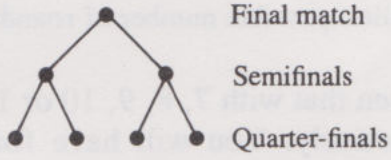


Figure 4.37

However, if there were only seven players we would have the situation shown in Figure 4.38; one of the players would not compete until the semifinals, and this player's bye in the first round is represented in Figure 4.38 by a white vertex. In total there are $3 + 2 + 1 = 6$ matches in this case.

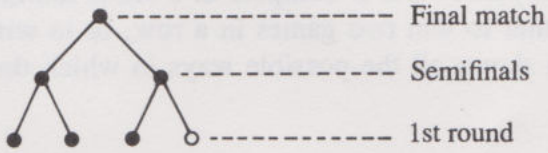


Figure 4.38

The tree for nine players is shown in Figure 4.39(a) – here there is only one first-round match, then four second-round, two third-round and a final, giving eight matches in total.

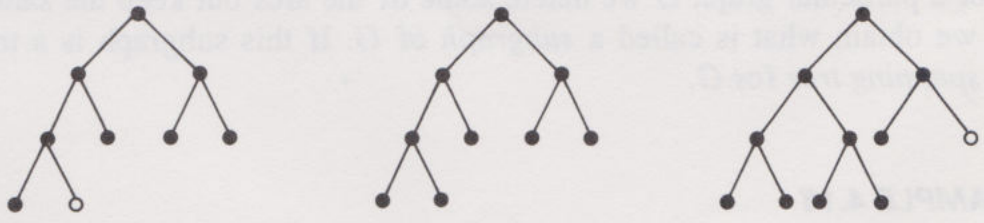


Figure 4.39

Trees representing tournaments with 10 and 11 players are given in Figure 4.39(b) and (c) respectively.

EXERCISE 4.24 For the tennis tournament in Example 4.17 draw a tree and hence determine the smallest number of matches when there are (a) 15, (b) 19, (c) 27 competitors.

EXERCISE 4.25 A total of six black and six white balls is distributed between three boxes. Box 1 contains one black and two white balls, box 2 contains two black and three white balls, and box 3 contains three black and one white balls. You select a container at random and take one ball out of it (also at random). Draw a tree which represents all the possible outcomes. Indicate on each arc the probability that the particular event occurs. Hence determine whether it is more likely that a black ball or a white ball is chosen.

EXERCISE 4.26 Twenty teams have entered for a 'knockout' football competition, in which the winners of round 1 proceed to round 2, and so on, until a single winning team emerges. There are no drawn games. Draw a tree showing a scheme for the tournament, with the smallest possible number of rounds and all byes in round 1.

In Example 4.17 it was seen that with 7, 8, 9, 10 or 11 players there would be 6, 7, 8, 9 or 10 matches respectively. You will have found similar results for the 'knockout competition' problems in Exercises 4.24 and 4.26. In fact, it's easy to generalize this: if there are n players in a knockout tournament then there will be exactly $n-1$ matches. This is because every player loses exactly one match (and is then out of the competition), with the exception of the overall winner who does not lose any match.

EXERCISE 4.27 Two players A and B compete in a chess tournament. The winner is the one who is first either to win two games in a row, or to win a total of three games. Draw a tree which shows all the possible ways in which the game can proceed to a conclusion.

EXERCISE 4.28 A frog hops along a straight line (the x-axis). It begins at the origin, and each jump has unit length either to the right or to the left. It stops either when it has made a total of four jumps, or if it reaches $x=3$ or $x=-2$. Draw a tree which represents all the possible paths the frog can travel.

If for a particular graph G we delete some of the arcs but keep the same set of vertices we obtain what is called a *subgraph* of G. If this subgraph is a tree, it is called a *spanning tree* for G.

■ *EXAMPLE 4.18*

For the graph in Figure 4.26, some (but not all) of the spanning trees are shown in Figure 4.40.

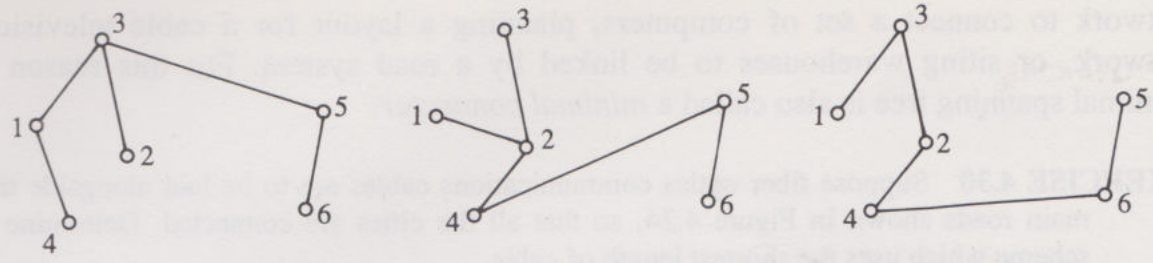


Figure 4.40

■ *EXAMPLE 4.19*

A spanning tree is shown in Figure 4.41(b) for the graph in Figure 4.41(a).

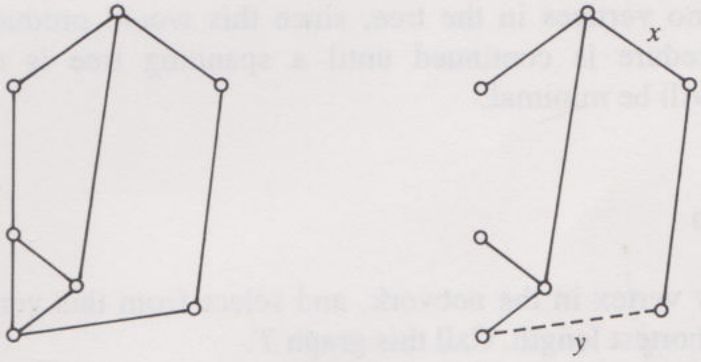


Figure 4.41

If the arc x is removed then the graph in (b) is no longer a spanning tree, since it becomes a disconnected graph. Alternatively, if the arc y (indicated by a dashed line) is added then (b) is no longer a tree since it contains a cycle.

EXERCISE 4.29 Draw the eight spanning trees for the graph in Figure 4.42.

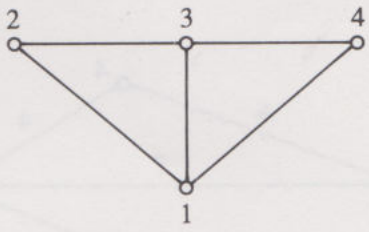


Figure 4.42

If we have a network where the weights on the arcs represent distance, then a *minimal spanning tree* is a spanning tree which has the least possible length. It is important to determine minimal spanning trees for applications such as designing a

network to connect a set of computers, planning a layout for a cable television network, or siting warehouses to be linked by a road system. For this reason a minimal spanning tree is also called a *minimal connector*.

EXERCISE 4.30 Suppose fiber optics communications cables are to be laid alongside the main roads shown in Figure 4.24, so that all the cities are connected. Determine a scheme which uses the shortest length of cable.

You shouldn't have had too much trouble solving Exercise 4.30 by trial and error – in fact, there are *two* possible minimal connectors. In a large problem, however, we need a systematic way of finding a minimal spanning tree, and this is provided by *Prim's method*. We begin constructing the tree with one arc. We then add one arc which is the shortest of the remaining arcs which have *one* vertex in the tree (we reject arcs which have two vertices in the tree, since they would produce a cycle, and arcs which have no vertices in the tree, since this would produce a disconnected graph). The procedure is continued until a spanning tree is obtained, and by construction this will be minimal.

Prim's Algorithm

Step 1 Take any vertex in the network, and select from this vertex the arc which has the shortest length. Call this graph T.

Step 2 Select the arc (i, j) with the smallest length from amongst *all* arcs (i, k) with i in T and k not in T. Add this arc to T.

Step 3 If T is a spanning tree for the given network, the solution has been obtained. Otherwise, repeat Step 2.

■ *EXAMPLE 4.20*

We'll use Prim's algorithm to find a minimal spanning tree for the network shown in Figure 4.43.

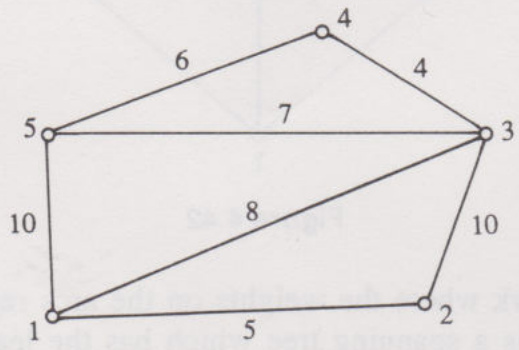


Figure 4.43

Step 1 Begin at vertex 1 and select arc (1, 2) which is the shortest arc originating at this vertex. Hence

$T = \{(1, 2)\}$

Step 2 Consider all the arcs which have one vertex in T:

(1, 5), (1, 3), (2, 3)

The shortest is (1, 3) so we add this to T, producing

$T = \{(1, 2), (1, 3)\}$

Step 3 T is not a spanning tree; go to

Step 2 Consider arcs with one vertex in T:

(1, 5), (3, 4), (3, 5)

The shortest is (3, 4); adding this to T gives

$T = \{(1, 2), (1, 3), (3, 4)\}$

Step 3 T is not a spanning tree; go to

Step 2 Consider arcs

(1, 5), (3, 5), (4, 5)

The shortest is (4, 5), giving

$T = \{(1, 2), (1, 3), (3, 4), (4, 5)\}$

Step 3 T is now a spanning tree, and therefore is a minimal spanning tree with length 23, shown in Figure 4.44.

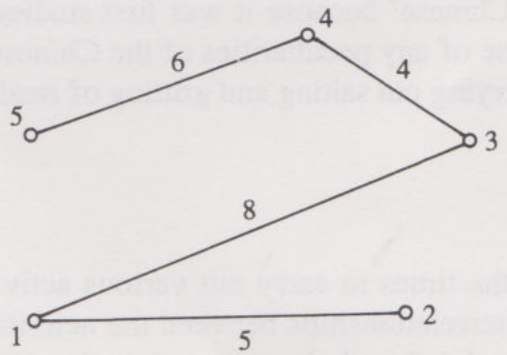


Figure 4.44

EXERCISE 4.31 Repeat Example 4.20, starting at vertex 3.

EXERCISE 4.32 Use Prim's algorithm to find a minimal spanning tree for the network in Figure 4.45 starting at vertex 7. Repeat, starting at vertex 1, to obtain a *different* minimal spanning tree.

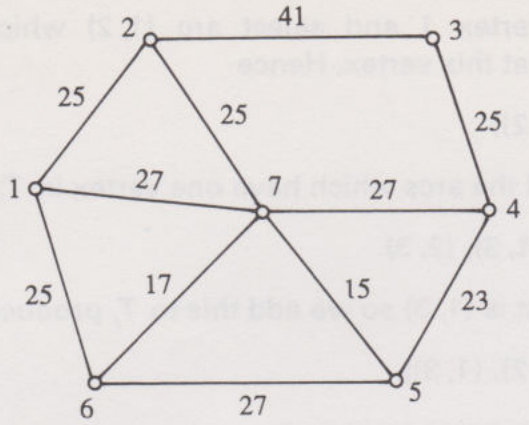


Figure 4.45

Three other optimization problems associated with networks are as follows.

The travelling salesperson problem

A salesperson has to visit a number of towns and return to the starting point. Each town is to be visited exactly once, and the travelling times between the towns are known. The problem is to carry out the tour in the shortest possible time.

The Chinese postperson problem

A postperson picks up mail at the sorting office and has to deliver it in a certain district, and then return to the depot. Each street in the district has to be covered at least once. The problem is to choose a route requiring the shortest possible distance to be travelled. This problem is called 'Chinese' because it was first studied by a Chinese mathematician in 1962, not because of any peculiarities of the Chinese postal system! A similar problem faces a truck carrying out salting and gritting of roads in winter.

Critical path analysis

For a complex project, the times to carry out various activities are represented on a network which shows interrelationships between the activities. For example, to build a house involves laying the foundations, erecting the walls, putting on the roof, installing the electric wiring and plumbing, plastering, decorating and many other jobs. The problem is to find a 'critical path' through the network which identifies the minimum time in which the project can be completed.

We don't have space to consider these problems here, but they are dealt with in several of the books listed at the end of this chapter. In particular, the book by Wilson and Watkins (1990) also describes other interesting applications of graphs.

We'll end this section with a brief look at how matrices can be applied to graphs. First, go back to the idea of the bipartite graph, an example of which was shown in Figure 4.25.

■ *EXAMPLE 4.21*

The bipartite network in Figure 4.46 represents connections between three airports A_1, A_2, A_3 in country A with airports B_1, B_2 in a second country B.

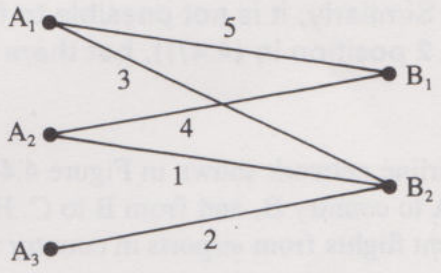


Figure 4.46

The numbers on the arcs are the numbers of different airlines flying on that route – for example, there are four airlines offering flights from A_2 to B_1. We can express the information provided by this network in the matrix

$$T=\begin{bmatrix}5 & 3\\ 4 & 1\\ 0 & 2\end{bmatrix}$$

The rows correspond to the airports A_1, A_2, A_3 and the columns to B_1 and B_2. Each entry in row i, column j gives the number of different flights from A_i to B_j, so for example the 2, 1 entry is 4. Suppose that there are onward connections to three airports C_1, C_2, C_3, in a third country C, shown in Figure 4.47.

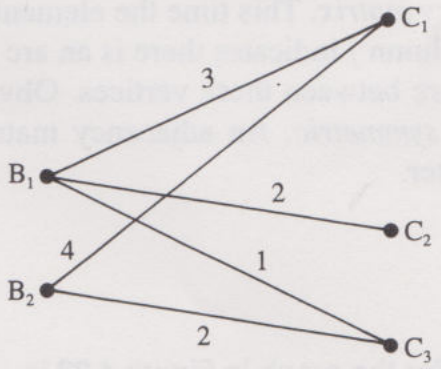


Figure 4.47

The matrix for this network is

$$S=\begin{bmatrix}3 & 2 & 1\\ 4 & 0 & 2\end{bmatrix}$$

where the rows correspond to airports in country B and columns to country C. If we combine together the two networks, then the matrix showing flights between countries A and C is just the product

$$TS = \begin{bmatrix} 27 & 10 & 11 \\ 16 & 8 & 6 \\ 8 & 0 & 4 \end{bmatrix} \tag{4.47}$$

This means, for example, that there are four ways of flying from A_3 to C_3 – you can check this by looking at Figures 4.46 and 4.47, where you'll see there are two ways of flying from A_3 to B_2, and two ways from B_2 to C_3, giving four different ways in total. Similarly, it is not possible to fly from A_3 to C_2 (shown by the zero entry in the 3, 2 position in (4.47)), but there are 10 ways of going from A_1 to C_2.

EXERCISE 4.33 For the airline network shown in Figure 4.48 write down the matrices for flights from country A to country B, and from B to C. Hence obtain the matrix giving the numbers of different flights from airports in country A to those in country C.

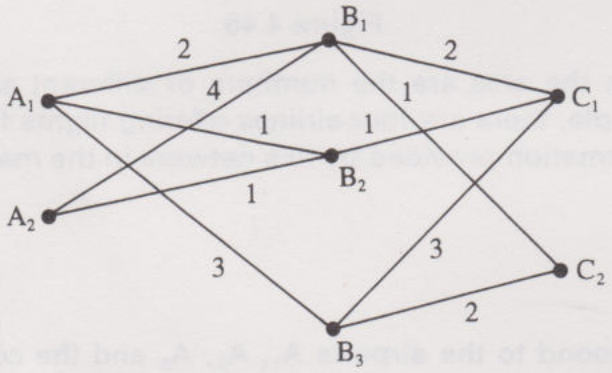


Figure 4.48

If we have a graph rather than a network then we can still associate a matrix A with it, called the *adjacency matrix*. This time the elements of A are either 1 or 0. An element $a_{ij} = 1$ in row i, column j indicates there is an arc connecting vertices i and j. If $a_{ij} = 0$ then there is no arc between these vertices. Obviously $a_{ij} = a_{ji}$ for all i and j, which means that A is *symmetric*. An adjacency matrix is a convenient way of storing a graph in a computer.

■ *EXAMPLE 4.22*

The adjacency matrix for the graph in Figure 4.23 is

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \end{bmatrix} \tag{4.48}$$

There are five vertices in this graph, so A is a 5×5 matrix. In general if there are n vertices then A is $n \times n$. Notice also that in (4.48) all the elements on the principal diagonal are zero. In general the i, i element of an adjacency matrix will be 1 only if there is an arc connecting vertex i to itself (such an arc is called a *loop*). You can see that A in (4.48) is symmetric: the elements in the first row are the same as those in the first column, the second row is the same as the second column, and so on.

The adjacency matrix can be used to investigate certain properties of graphs. Recall that a *walk* is a set of arcs joining one vertex to another in which repetitions of arcs are allowed – that is, we can 'retrace our steps'. Illustrations of walks were given in Example 4.12. In particular, the walk given in part (iii) of Example 4.12(b) is a walk from vertex 1 to vertex 3 for the graph in Figure 4.26 involving *four* arcs. It matters in which order the arcs are taken: for example, in Figure 4.26 the two walks

$$\{(1,2),(2,3),(3,1)\}, \quad \{(1,3),(3,2),(2,1)\}$$

from vertex 1 to itself are different from each other.

It can be shown that if we raise the adjacency matrix for a graph to the pth power then the element in row i, column j of A^p is the number of different walks from vertex i to vertex j using p arcs.

■ *EXAMPLE 4.23*

Consider the graph in Figure 4.42. Its adjacency matrix is

$$A = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}$$

Using the standard multiplication rule given in Section 1.4, Chapter 4, it's easy to multiply A by itself to get

$$B = A^2 = \begin{bmatrix} 3 & 1 & 2 & 1 \\ 1 & 2 & 1 & 2 \\ 2 & 1 & 3 & 1 \\ 1 & 2 & 1 & 2 \end{bmatrix} \tag{4.49}$$

The element b_{ij} in the i, j position of A^2 is the number of different walks between vertices i and j in Figure 4.42 using *two* arcs. For example, for $b_{11} = 3$, the three walks from vertex 1 to itself using two arcs are

(1, 2), (2, 1)
(1, 2), (3, 1)
(1, 4), (4, 1)

Notice that in each of these we retrace our steps. For $b_{12} = 1$, the only walk from vertex 1 to vertex 2 using two arcs is

(1, 3), (3, 2)

and for $b_{13} = 2$, the two walks from vertex 1 to vertex 3 are

(1, 2), (2, 3)
(1, 4), (4, 3)

Notice that A^2 in (4.49) is also symmetric, so only the elements on and above the principal diagonal (top left corner to bottom right corner) need to be computed. In fact, the portion of A^2 below the principal diagonal doesn't give us any extra information: for example, $b_{32} = b_{23}$, which simply says that the number of walks from vertex 3 to vertex 2 using two arcs is the same as the number from vertex 2 to vertex 3.

EXERCISE 4.34 List the walks corresponding to all the remaining elements of A^2 in (4.49).

EXERCISE 4.35 Multiply A^2 in (4.49) by A to obtain A^3. List the walks using three arcs, corresponding to the elements of A^3.

EXERCISE 4.36 Using the adjacency matrix A in (4.48) for the graph in Figure 4.23, compute A^2. List the walks corresponding to its elements.

4.5 OPTIMAL CONTROL

In the discussion at the end of Example 3.2 in Chapter 3 we briefly introduced the notion of optimal control. The basic idea is to control a system in some 'best possible way' according to a particular aim or objective. The simple model discussed in Example 3.2 was of a car being driven along a straight road; a suggested optimization problem was to drive from one set of green traffic lights to the next set at red, either in the shortest possible time, or using the least possible amount of fuel. Clearly the control which performs the required task (i.e. the 'optimal' control) will be quite different according to which objective is aimed at. You probably feel instinctively that in the first case, to do the journey as quickly as possible you would accelerate flat out, and then put your foot hard down on the brake pedal so as to screech to a stop at the red light! In the second case you would minimize the fuel consumption by accelerating gently, and then coasting along before coming to a standstill at the light.

Finding an optimal control even for a simple system involves quite a bit of calculus, so you may wish to skip this section if you have little or no experience in differentiation and integration. As usual, we'll try and keep the technicalities to a minimum, but this last section of the book is something of a bridge to more advanced work on optimization.

■ *EXAMPLE 4.24*

The very simplest control model is described by a single equation

$$\frac{dx}{dt} = u \tag{4.50}$$

where the state variable $x(t)$ and the control variable u are both functions of time t. Suppose that initially at time $t = 0$ the system is at the origin, that is

$$x(0) = 0 \tag{4.51}$$

and it is required to choose u so as to transfer the system to $x = 1$ at $t = 1$, that is

$$x(1) = 1 \tag{4.52}$$

There are many control functions which will perform this transfer – indeed, as many as you like! Some simple functions which do the trick are

$$\begin{aligned} u &= 1 \\ u &= 2t \\ u &= 4 - 6t \end{aligned} \tag{4.53}$$

If we solve the differential equation (4.50) in each case, simply by integrating each of the expressions in (4.53) and using the initial condition (4.51), the three corresponding solutions $x(t)$ of (4.50) are

$$\begin{aligned} x &= t \\ x &= t^2 \\ x &= 4t - 3t^2 \end{aligned} \tag{4.54}$$

You can easily check that each of the expressions in (4.54) satisfies the conditions (4.51) and (4.52), and when differentiated gives the corresponding term in (4.53). The graphs of the functions in (4.54) are shown in Figure 4.49.

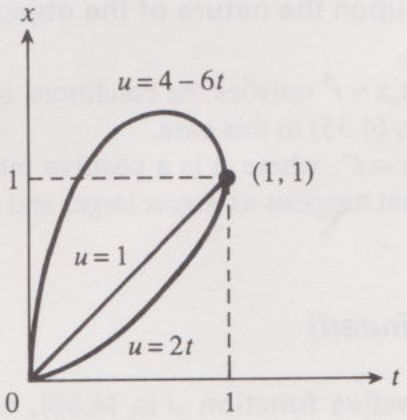


Figure 4.49

Clearly there are an infinite number of curves connecting the origin 0 with the point (1, 1). The idea of optimal control is to select a control which minimizes some 'objective function'. Suppose in this example we think of u as some kind

of control force, and we want to minimize the energy used in transferring the system from its initial state to its final state. The energy used is proportional to u^2, because energy is required whether u is positive or negative. The total energy consumed from $t=0$ to $t=1$ is

$$J=\int_0^1 u^2\,dt \tag{4.55}$$

and J is our *objective function*. If we substitute into (4.55) the three different controls (4.53) we get, using standard results in integration,

$$\int_0^1 1^2\,dt=[t]_0^1=1$$

$$\int_0^1 (2t)^2\,dt=[\tfrac{4}{3}t^3]_0^1=\tfrac{4}{3}$$

$$\int_0^1 (4-6t)^2\,dt=\left[\frac{(4-6t)^3}{-18}\right]_0^1=4$$

The first control $u=1$ in (4.53) therefore gives the smallest value of J in (4.55), but we can't say that it is optimal because we've only tried three out of the infinite number of possible controls.

We see in Figure 4.49 that the third control $u=4-6t$ in (4.53) causes the system to 'overshoot' its target – that is, it goes past $x=1$ and then returns. If we want to avoid this we might decide to minimize the area under the curve $x(t)$, that is take instead of (4.55) a different objective function

$$J=\int_0^1 x\,dt \tag{4.56}$$

In fact we can see from Figure 4.49, without doing any calculations, that the area under the curve is smallest when $u=2t$ than for the other two cases. Again, we can't be sure that $u=2t$ is the best possible control to minimize J in (4.56), but we have illustrated the fact that the selection of an optimal control will depend very much upon the nature of the objective function.

EXERCISE 4.37 Verify that $x=t^3$ satisfies the conditions (4.51) and (4.52). Obtain u from (4.50) and evaluate J in (4.55) in this case.

Repeat this with $x=t^n$, where n is a positive integer, and obtain the expression for J in terms of n. What happens as n gets larger and larger?

■ *EXAMPLE 4.24 (continued)*

When we used the objective function J in (4.55), of the three controls in (4.53) we found that $u=1$ was the best, since it gave the smallest value of J. We now establish that this is the best of *all* possible controls, by introducing the idea of the *hamiltonian function* (named after the nineteenth-century Irish mathematician Hamilton) which is defined by

$$H=pu+u^2 \tag{4.57}$$

In (4.57) p is a *new* variable, which also depends upon time t, and is called the *adjoint variable*. The expression (4.57) is obtained by multiplying the right-hand side of the differential equation (4.50) by p, and adding the quantity u^2 from inside the integral (4.55). It can be shown that the optimal control u^* is given by the condition

$$\frac{\partial H}{\partial u} = 0 \tag{4.58}$$

where this notation, called the *partial derivative* of H with respect to u, simply means that we differentiate H in the usual way, *regarding everything except u* as a constant. Applying (4.58) to (4.57), in which we regard p as a constant so far as differentiation is concerned, gives us

$$p + 2u = 0$$

so that the optimal control is

$$u^* = -\tfrac{1}{2}p \tag{4.59}$$

We appear to be no further forward, since we don't know what this mysterious 'adjoint variable' is! In fact, p satisfies another differential equation:

$$\frac{dp}{dt} = -\frac{\partial H}{\partial x} \tag{4.60}$$

where in (4.60) the notation $\partial H/\partial x$ means the partial derivative of H with respect to x, regarding everything except x as a constant. Indeed, since H in (4.57) doesn't contain x at all, we have $\partial H/\partial x = 0$, so (4.60) becomes

$$\frac{dp}{dt} = 0$$

From this we deduce that $p = a$, where a is a constant. We're now getting nearer, since we can say from (4.59) that the optimal control is

$$u^* = -\tfrac{1}{2}a$$

It remains to find the value of the constant a, and to do this we use the conditions $x(0) = 0$, $x(1) = 1$. The original differential equation (4.50) is now

$$\frac{dx}{dt} = u^*$$

$$= -\tfrac{1}{2}a$$

and integrating this with respect to t gives

$$x = -\tfrac{1}{2}at + b$$

where b is another 'constant of integration'. However, since $x = 0$ when $t = 0$ it follows that $b = 0$, so that x reduces to

$$x = -\tfrac{1}{2}at$$

Finally, since $x = 1$ when $t = 1$ we must have $a = -2$, so that the optimal control is indeed

$$u^* = -\tfrac{1}{2}a = 1$$

as we suggested above.

■ EXAMPLE 4.25

Let's look at a slightly more complicated case where the system is described by the single equation

$$\frac{dx}{dt} = x + u \tag{4.61}$$

instead of (4.50), but the objective function J to be minimized remains the expression in (4.55), and the initial and final conditions (4.51) and (4.52) are also the same. As before we introduce an adjoint variable p and multiply it by the right-hand side of (4.61), and add on u^2 to produce the hamiltonian

$$\begin{aligned} H &= p(x + u) + u^2 \\ &= px + pu + u^2 \end{aligned} \tag{4.62}$$

The optimal control u^* is given by (4.58). However, the terms involving u in (4.62) are the same as those in the previous example in (4.57), so as before we have

$$\frac{\partial H}{\partial u} = p + 2u$$

and setting this equal to zero as in (4.58) gives the optimal control

$$u^* = -\tfrac{1}{2}p$$

However, this time H in (4.62) contains a term px, so the partial derivative $\partial H/\partial x$ is equal to p. Equation (4.60) therefore becomes

$$\begin{aligned} \frac{dp}{dt} &= -\frac{\partial H}{\partial x} \\ &= -p \end{aligned}$$

The solution of this equation is

$$p = ae^{-t} \tag{4.63}$$

where a is a constant of integration, as you can verify by differentiation (we discussed equations like (4.63) in Exercise 3.18 in Chapter 3). We therefore have

$$\begin{aligned} u^* &= -\tfrac{1}{2}p \\ &= -\tfrac{1}{2}ae^{-t} \end{aligned} \tag{4.64}$$

Substituting the expression (4.64) into the state equation (4.61) gives us

$$\frac{dx}{dt} = x - \tfrac{1}{2}ae^{-t} \tag{4.65}$$

This isn't the place to go into solving differential equations like (4.65) – there are very many books where this is covered. We'll simply state that the solution of (4.65) is

$$x = be^{t} + \tfrac{1}{4}ae^{-t} \tag{4.66}$$

where b is a constant of integration, and leave it to you to verify, if you wish, that when (4.66) is differentiated it satisfies (4.65). The values of the

constants a and b in (4.66) are found by using the conditions at $t=0$ and $t=1$. These give

$$x(0)=0: \quad b+\tfrac{1}{4}a=0$$

$$x(1)=1: \quad be+\tfrac{1}{4}ae^{-1}=1$$

and some simple algebra to solve these simultaneous equations produces

$$a=\frac{-4}{e-e^{-1}}$$

Hence from (4.64) we obtain the optimal control to be

$$u^*=\frac{2e^{-t}}{e-e^{-1}}$$

EXERCISE 4.38 Repeat Example 4.25 if the system is to be transferred from $x(0)=2$ to $x(1)=0$.

EXERCISE 4.39 A system is described by the equation

$$\frac{dx}{dt}=-x+u$$

The control u is to be chosen so as to minimize the objective function

$$\int_0^1 u^2\,dt$$

whilst transferring the system from $x(0)=1$ to $x(1)=0$. Show that the optimal control is

$$u^*=\frac{-2e^t}{e^2-1}$$

So far we have only considered control systems described by a single differential equation. As we saw in Chapter 3, realistic models of systems will involve many state variables. Naturally, the mathematics required to find an optimal control becomes more complicated, but the basic step of constructing a hamiltonian function remains unaltered.

■ *EXAMPLE 4.26*

Consider a carriage which runs along smooth straight rails, and has a rocket motor at each end, as shown in Figure 4.50. This can perhaps be thought of as

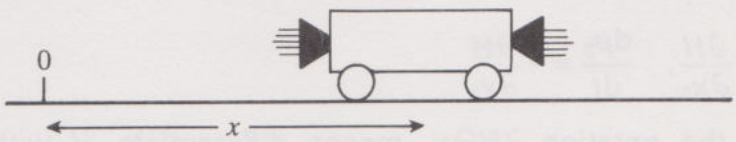


Figure 4.50

the ultimate high-speed train, hurtling through the Channel Tunnel! According to Newton's law of motion given in equation (3.3) in Chapter 3, if we assume for simplicity that the carriage has unit mass then the equation of motion is just

$$\frac{d^2x}{dt^2} = u \tag{4.67}$$

where u is the net force exerted by the rocket motors. The problem is to transfer the carriage from rest at its starting point (the origin 0) to rest at its final destination (taken for convenience to be $x=1$ at time $t=1$), whilst minimizing the total energy required, namely

$$\int_0^1 u^2 \, dt \tag{4.68}$$

which is once again the same expression as in (4.55). We first define as the state variables the position x and velocity dx/dt of the carriage, that is

$$x_1 = x, \quad x_2 = \frac{dx}{dt} \tag{4.69}$$

so that the relationship between these variables is

$$\frac{dx_1}{dt} = x_2 \tag{4.70}$$

Also, the equation of motion (4.67) can now be written as

$$\frac{dx_2}{dt} = u \tag{4.71}$$

since $dx_2/dt = d^2x/dt^2$. We now have *two* differential equations (4.70) and (4.71) describing the state of the system. The definition of the hamiltonian function now requires *two* adjoint variables, one multiplying the right-hand side of (4.70) and the other the right-hand side of (4.71), so we get

$$H = p_1 x_2 + p_2 u + u^2 \tag{4.72}$$

where as before we add on the quantity u^2 from inside the integral (4.68) to be minimized. The optimal control u^* is still given by setting $\partial H/\partial u = 0$ as in (4.58), where in the differentiation of (4.72) we regard as constant all the variables except u. This gives us

$$p_2 + 2u = 0$$

so that

$$u^* = -\tfrac{1}{2}p_2 \tag{4.73}$$

As in (4.60), each adjoint variable satisfies a differential equation, and these are now

$$\frac{dp_1}{dt} = -\frac{\partial H}{\partial x_1}, \quad \frac{dp_2}{dt} = -\frac{\partial H}{\partial x_2} \tag{4.74}$$

As before, the notation $\partial H/\partial x_1$ means differentiate H with respect to x_1, regarding everything else as constant; similarly for $\partial H/\partial x_2$. Applying (4.74) to H

in (4.72) produces

$$\frac{dp_1}{dt} = 0, \quad \frac{dp_2}{dt} = -p_1 \tag{4.75}$$

since H does not contain x_1, and the term in x_2 is $p_1 x_2$. It follows from the first equation in (4.75) that $p_1 = a$, where a is a constant. Hence from the second equation in (4.75) we have

$$\frac{dp_2}{dt} = -a$$

which gives

$$p_2 = -at + b$$

where b is another constant of integration. The optimal control is therefore given by (4.73) as

$$u^* = \tfrac{1}{2}at - \tfrac{1}{2}b \tag{4.76}$$

and substituting this into (4.71) gives

$$\frac{dx_2}{dt} = \tfrac{1}{2}at - \tfrac{1}{2}b$$

Integrating with respect to t produces

$$x_2 = \tfrac{1}{4}at^2 - \tfrac{1}{2}bt + c \tag{4.77}$$

where c is another constant. We can now put (4.77) into (4.70) and integrate yet again with respect to t to obtain

$$x_1 = \tfrac{1}{12}at^3 - \tfrac{1}{4}bt^2 + ct + d \tag{4.78}$$

where d is a further constant. We now use the conditions at $t = 0$ and $t = 1$ to obtain the values of the constants a, b, c, d. Since the system starts at the origin with zero velocity x_2 we have

$$t = 0, \quad x_1 = 0, \quad x_2 = 0$$

which when substituted into (4.77) and (4.78) gives $c = 0$, $d = 0$. At the end of the journey when $x_1 = 1$ and again the velocity x_2 is zero we have

$$t = 1, \quad x_1 = 1, \quad x_2 = 0$$

which when substituted into (4.77) and (4.78) gives

$$\tfrac{1}{4}a - \tfrac{1}{2}b = 0$$

$$\tfrac{1}{12}a - \tfrac{1}{4}b = 1$$

Solving these equations gives $a = -24$, $b = -12$, so finally from (4.76) the optimal control is

$$u^* = -12t + 6 \tag{4.79}$$

Notice that this optimal control in (4.79) starts off at $u^* = 6$ units, decreases to zero at $t = \frac{1}{2}$ and ends up at -6 units. This is what we would expect: referring to

Figure 4.50, in the first half of the journey the right-hand rocket is switched off, the left-hand rocket accelerates the carriage by starting off at 6 units and steadily reducing to zero; in the second half of the trip the left-hand rocket is off, the right-hand rocket decelerates the carriage by steadily increasing its thrust from zero to 6 units (the negative sign occurs in u^* because this thrust is in the negative x-direction).

EXERCISE 4.40 With the optimal control in (4.79) determine the value of the objective function in (4.68).

Our discussion of controllability in Section 3.2 of Chapter 3 assumed that there were no restrictions on the magnitudes of the control variables, but in practice these are very often present. To get some idea of what happens in such cases, let's continue with our rocket-propelled carriage which is still to make the same trip from rest at $x=0$ to rest at $x=1$. However, suppose now that the rockets can exert a maximum thrust of 1 unit. Since each can fire in only one direction, the net thrust can be at most 1 unit to the left or right, that is $-1 \leqslant u \leqslant 1$.

The control (4.79) could then not be used, since it required thrusts greater than 1 unit. To simplify matters, we'll consider minimizing the total *time* T of the journey, instead of minimizing energy used. It turns out that the optimal strategy in this case is

$$\begin{aligned} u^* &= +1, \quad 0 \leqslant t \leqslant \tfrac{1}{2}T \\ u^* &= -1, \quad \tfrac{1}{2}T < t \leqslant T \end{aligned} \tag{4.80}$$

This means that for the first half of the trip the left-hand rocket is used at maximum thrust, giving maximum possible acceleration, and for the second half of the trip the right-hand rocket is also used at maximum thrust, giving maximum possible deceleration. Thus no intermediate values of rocket thrust are used – a motor is either full on, or off. Because of this property, the optimal control (4.80) is given the graphic name 'bang–bang control' – a rare example of informal language being used as a technical term! In fact the concept of bang–bang control clarifies our intuitive discussion at the beginning of this section on how to drive a car from one set of traffic lights to another as quickly as possible: 'bang' the foot down on the accelerator pedal, and then on the brake! However, investigation of how we derive bang–bang controls is well beyond the mathematical level of this book. Instead, we end by giving a couple of examples of interesting optimal control problems, without attempting to solve them.

■ *EXAMPLE 4.27*

A spacecraft is a distance h from an asteroid whose mass is so small that its gravitational attraction can be ignored. A small landing vehicle separates

from the spacecraft at $t=0$ with initial downwards velocity v, and the objective is to achieve a 'soft landing' on the asteroid – that is, at the instant of touchdown (time $t=T$) the vertical velocity is to be zero. Only 'vertical' motion is considered. Let x_1 be the altitude and x_2 the velocity of the lander at time t. The equations of motion are just (4.70) and (4.71). Suppose that it is required to minimize a combination of fuel used and the time T to landing. We also suppose that the rocket motor can fire either up or down, with maximum magnitude 1 unit. The total fuel consumption can be represented as

$$C=\int_0^T |u|\,dt$$

where $|u|$ denotes the magnitude of the control, that is ignoring its sign, since fuel is used whatever the direction of the rocket thrust. The overall objective function to be minimized is $C+T$. It can be shown that provided $v^2/2 < h < 5v^2/2$ then the optimal control in this case has 'zero–bang' form, that is

$$u^* = 0, \quad 0 \leqslant t \leqslant t_1$$

$$u^* = 1, \quad t_1 < t \leqslant T$$

where $t_1 = h/v - v/2$, and the minimum time to landing is

$$T=\frac{h}{v}+\frac{v}{2}$$

Thus the vehicle coasts down at constant velocity v with the motor off, until at time t_1 the motor is switched on to give maximum possible upwards thrust, so that on reaching the asteroid's surface the vehicle touches down with zero velocity.

■ *EXAMPLE 4.28 Cash balance model*

A firm has a known demand for cash over a period of time T. In order to meet this demand the firm must have access to funds, in the form of either actual cash or investments. There are two conflicting aspects of this: if the firm holds too much cash then it loses money which could be earned by investments; alternatively, if too little cash is held then the firm has to sell investments to meet the cash demand, and thereby incurs a broker's commission. The problem is to find an optimal trade-off between the amounts kept in cash and investments.

Let x_1 and x_2 be the balances held in cash and investments at time t, and let the known demand be d, which will vary with time. The control variable u is the amount of investments bought or sold, where a negative value means a purchase. In practice u will be subject to limits on the amounts which can be bought or sold, but a cost $\alpha|u|$ is incurred for each transaction, where α is the broker's commission rate. Let r_1 and r_2 be the rates of interest on the

cash and investments respectively. The equations describing the state of the system are

$$\frac{dx_1}{dt} = r_1 x_1 - d + u - \alpha |u|$$

rate of increase of cash balance | interest earned on cash | cash paid out | cash from investments sold | broker's commission

$$\frac{dx_2}{dt} = r_2 x_2 - u$$

rate of increase of investment balance | interest earned on investments | investments sold for cash

Starting with known balances $x_1(0)$, $x_2(0)$, the problem is to determine the control which maximizes the net sum $x_1(T) + x_2(T)$ held by the firm at the end of the period under consideration. The optimal control turns out to have 'bang–zero–bang' form, where investments are either bought or sold in the maximum allowable amounts, or not at all.

EXERCISE 4.41 A system is described by the equation

$$\frac{dx}{dt} = u$$

where the control u is to be chosen so as to minimize the objective function

$$\int_0^1 (x^2 + u^2) dt$$

whilst transferring the system from $x(0) = 0$ to $x(1) = 1$. Construct the hamiltonian, and show that under the influence of the optimal control u^* the system satisfies the equation

$$\frac{d^2x}{dt^2} = x$$

Verify that the solution of this equation is

$$x = \frac{1}{e - e^{-1}} (e^t - e^{-t})$$

and show that

$$u^* = \frac{1}{e - e^{-1}} (e^t + e^{-t})$$

EXERCISE 4.42 An inventory-control production-scheduling problem is described by the equation

$$\frac{dI}{dt} = P$$

where I is the inventory level and P is the production rate, after known sales demand has been met. It is required to control the production rate P from time $t=0$ to $t=T$ so as to minimize the objective function

$$\int_0^T (I^2 + P^2)\mathrm{d}t$$

Set up the hamiltonian function, and show that the adjoint variable p satisfies the equation

$$\frac{\mathrm{d}^2 p}{\mathrm{d}t^2} = p$$

Hence show that the optimal control has the form

$$P^* = ae^t + be^{-t}$$

where a and b are constants.

PROBLEMS

4.1 Use Fibonacci search to determine the value of x which maximizes the volume of the box in Exercise 4.2 to within an interval of uncertainty of length not more than 0.4 cm.

4.2 Use Fibonacci search with eight function evaluations to obtain the approximate value of x which maximizes

$$f(x) = \frac{4x-3}{1+x^2}, \quad 1 \leqslant x \leqslant 6.5$$

4.3 Use Fibonacci search with eight function evaluations to estimate the optimal radius of the beer can in Example 4.1, assuming $0 \leqslant x \leqslant 5$. Repeat with $2 \leqslant x \leqslant 5$ and seven function evaluations.

If you are familiar with the calculus approach, use it to obtain a more accurate solution.

Measure the radius of an actual 440 ml can. Why do you think it differs from the 'optimal' value you have calculated?

4.4 A tourist bus company has 19 minibuses each of which can carry 18 passengers, and 17 larger coaches which can carry 35 passengers. The company employs 30 drivers and 35 guides. Each bus carries only a single driver. The minibuses require only a single guide each, but two guides are carried on each of the larger buses. It is required to carry as many passengers as possible at any one time. Let x_1 denote the number of minibuses and x_2 the number of larger buses to be used. Express the problem in LP form. Sketch the feasible region, and find the optimal solution graphically and algebraically.

4.5 A tyre manufacturer operates two factories. Factory A produces 100 Super Tyres, 300 Excellent Tyres and 500 Budget Tyres per day. Factory B produces 200 of each kind

of tyre per day. The manufacturer has an order for 8000 Super, 16 000 Excellent and 20 000 Budget Tyres. The daily running costs for each factory are £20 000. The problem is to determine how many days each factor should be operated to fulfil the order as cheaply as possible.

Let x_1 and x_2 be the numbers of days factories A and B are open. Express the problem in LP form and solve graphically and algebraically.

4.6 A refinery makes three grades of petrol (P_1, P_2, P_3) from three crude oils (c_1, c_2, c_3). Crude type c_3 can be used in any grade, but the others must satisfy the following specifications:

Grade of petrol	Specification	Selling price (pence per litre)
P_1	Not less than 45% c_1 Not more than 25% c_2	65.3
P_2	Not less than 25% c_1 Not more than 60% c_2	53.5
P_3	No restrictions	52.1

There are capacity limits on the availabilities of the three crude oils, as follows:

Crude	Available capacity (kilolitres)	Cost (pence per litre)
c_1	150	61.2
c_2	160	50.8
c_3	80	54.9

Let x_{ij} litres be the amount of crude oil c_i used to make petrol P_j. Write down the constraints and the profit function which is to be maximized in the form of an LP problem.

4.7 Find an optimal solution to the transportation problem having the following table of availabilities, requirements and costs:

		Availabilities			
		15	20	30	35
Requirements	25	10	5	6	7
	25	8	2	7	6
	50	9	3	4	8

4.8 Consider the LP problem:

$$\text{Minimize} \quad z = 3x_1 + 2x_2 + 4x_4$$

$$\text{subject to:} \quad x_1 + 3x_2 + 8x_3 = 4$$

$$x_2 + 12x_3 + x_4 = 5$$

$$x_1, x_2, x_3, x_4 \geqslant 0$$

Starting with the feasible solution $x_1 = 4$, $x_2 = 0$, $x_3 = 0$, $x_4 = 5$ use the simplex technique to obtain an optimal solution.

4.9 An airline operates three types of aircraft on three different routes. The numbers of passengers (in thousands) which can be carried annually by each aircraft type are as follows:

		Aircraft type		
		1	2	3
Route	1	20	17	–
	2	–	18	15
	3	19	18	17

The available numbers of aircraft of each type are 15, 21 and 20 respectively. The costs per aircraft per year are (in certain units) as follows:

		Type		
		1	2	3
Route	1	17	19	–
	2	–	21	20
	3	16	15	14

The estimated numbers of passengers per year to be carried are as follows:

		Numbers (thousands)	Income per 1000 passengers
Route	1	290	15
	2	200	17
	3	230	10

It is required to allocate aircraft to routes so as to minimize the annual operating cost. If demand exceeds capacity, the cost of a 'lost' passenger is the amount of income lost.

Let

x_{ij} = number of aircraft on route i of type j

x_i = number of passengers (in thousands) who cannot be accommodated on route i

Express this problem in LP form.

4.10 In a horse-jumping competition a team has four horses and four riders. From past experience it is known how many penalty points rider i is likely to incur when riding horse j. Set this up as an assignment problem with the aim of matching riders to horses so as to minimize the total of expected penalty points.

4.11 A laboratory contains 25 microcomputers which must be connected to a power supply having four sockets. Connections are made using extension leads which each have four power sockets. Draw a tree which shows the minimum number of extension leads needed so that all the computers have power.

4.12 A box contains four black, five white and eight red balls. Two balls are taken out, one at a time, without replacement. Draw a tree representing all the possible outcomes. Label each arc with the probability that the appropriate event occurs.

What is the probability that one of the balls is red and the other is black?

4.13 A gambler decides to play at most five games of roulette. At each play the gambler either wins or loses £10. The gambler will stop before playing five games if he or she either goes broke, or wins a total of £30. The gambler begins with £10. Draw a tree showing all the possible outcomes.

In how many of these would the gambler withdraw before playing five games?

4.14 A graph G consisting of m individual trees is called a *forest*. Show that if G has an overall total of n vertices then it has a total of $n - m$ arcs.

Hence deduce that if a graph has n vertices, $n - 1$ arcs and contains no cycles then it is a tree.

4.15 The network in Figure 4.51 shows road distances in kilometres between certain towns and cities. A communications company wishes to connect together all these places by laying cables alongside these main roads. Find a layout which uses the least cable.

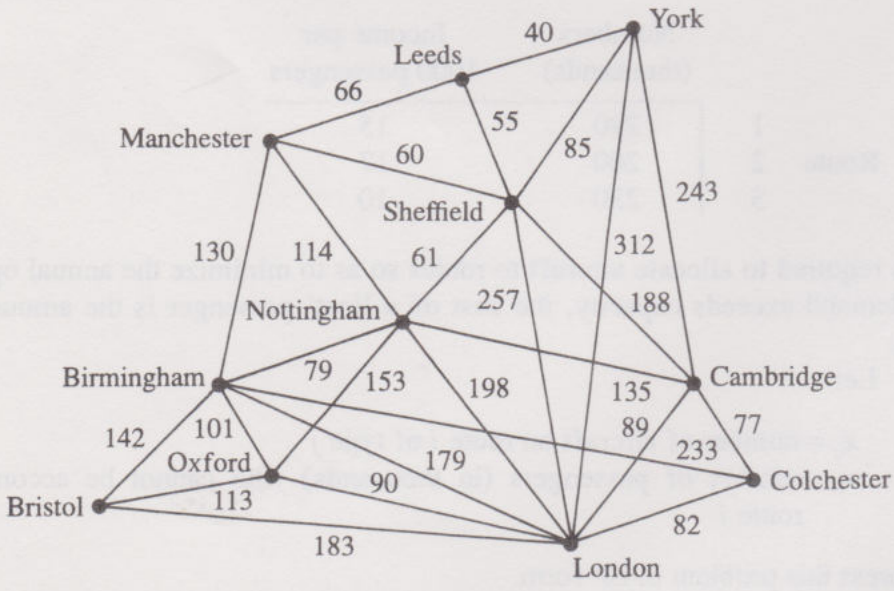


Figure 4.51

4.16 Consider the graph in Figure 4.52. The (3, 4) and (4, 3) elements of its adjacency matrix A are each equal to 2, because there are two arcs connecting vertices 3 and 4. The (1, 1) element is equal to 1 because of the arc connecting vertex 1 to itself. Write down A and determine A^2. List the walks involving two arcs corresponding to the elements of A^2.

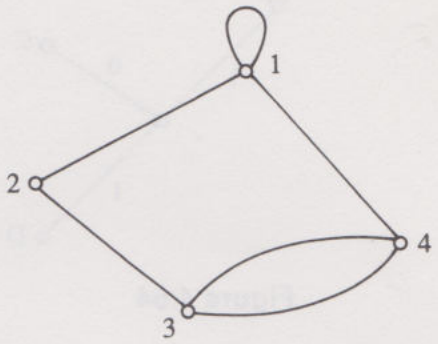


Figure 4.52

4.17 If a graph G has n vertices then to test whether it is connected, construct the adjacency matrix A and form the sum

$$S = A + A^2 + A^3 + \cdots + A^{n-1}$$

It can be shown that G is connected if and only if S has no zero elements.

Apply this procedure to the graph in Figure 4.53.

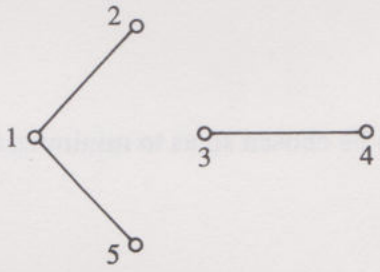


Figure 4.53

4.18 Trees can be used to decode messages where codewords have *variable* lengths. As a simple example consider the code

A	B	C	D
0	10	110	111

This can be represented by the tree in Figure 4.54.

Any received string of binary digits (assumed error free) can be decoded uniquely. This is because no codeword is the start of any other codeword.

Starting from vertex S, simply trace paths on the tree: for example, to decode 010111, the initial 0 takes you to A; returning to S, the bits 10 take you to B; finally, 111 takes you from S to D, so the message is decoded as ABD.

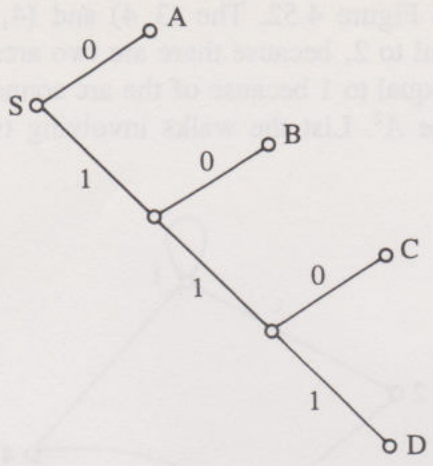


Figure 4.54

Construct the tree for the code

A	B	C	D	E
00	01	10	110	111

Decode the received messages

0001111110, 0000111101011001

4.19 A control system is described by the single equation

$$\frac{dx}{dt} = -2x + 2u \tag{4.81}$$

where the control u is to be chosen so as to minimize the objective function

$$\int_0^1 (3x^2 + u^2)dt$$

whilst transferring the system from $x(0) = 0$ to $x(1) = 1$. Show that under the influence of the optimal control u^* the state variable satisfies the equation

$$\frac{d^2x}{dt^2} = 16x$$

Hence show that in this case

$$x = \frac{e^{4t} - e^{-4t}}{e^4 - e^{-4}}$$

and deduce from (4.81) that

$$u^* = \frac{3e^{4t} + e^{-4t}}{e^4 - e^{-4}}$$

4.20 A control system is described by the equations

$$\frac{dx_1}{dt} = 3x_2$$

$$\frac{dx_2}{dt} = -2x_1 + 5x_2 + u$$

The control u is to be chosen so as to transfer the system from some given initial state to some given final state at time $t = 1$ whilst minimizing the objective function

$$\int_0^1 u^2 \, dt$$

Set up the hamiltonian function involving *two* adjoint variables p_1 and p_2. Show that $u^* = -\frac{1}{2}p_2$ where

$$\frac{d^2p_2}{dt^2} + 5\frac{dp_2}{dt} + 6p_2 = 0$$

Verify that this differential equation has solution

$$p_2 = c_1e^{-2t} + c_2e^{-3t}$$

where c_1 and c_2 are constants.

FURTHER READING

BARNETT, S. and CAMERON, R.G. 1985. *Introduction to Mathematical Control Theory*, 2nd Edition. Oxford University Press, Oxford.

BERRY, J., BURGHES, D. and HUNTLEY, I. (Eds). 1990. *Decision Mathematics*. Ellis Horwood, Chichester.

BIGGS, N.L. 1985. *Discrete Mathematics*. Oxford University Press, Oxford.

BONDI, C. (Ed.). 1991. *New Applications of Mathematics*. Penguin, London.

BURGHES, D.N. and DOWNS, A.M. 1975. *Modern Introduction to Classical Mechanics and Control*. Ellis Horwood, Chichester.

BURGHES, D.N. and GRAHAM, A. 1980. *Introduction to Control Theory, including Optimal Control*. Ellis Horwood, Chichester.

CHARTRAND, G. 1977. *Introductory Graph Theory*. Dover, New York.

COMPTON, C. and RIGBY, G. 1992. *Discrete and Decision Mathematics*. Hodder and Stoughton, Sevenoaks, Kent.

DIERKER, P.F. and VOXMAN, W.L. 1986. *Discrete Mathematics*. Harcourt Brace Jovanovich, San Diego.

FRENCH, S., HARTLEY, R., THOMAS, L.C. and WHITE, D.J. 1986. *Operational Research Techniques*. Edward Arnold, London.

GASS, S.I. 1990. *An Illustrated Guide to Linear Programming*. Dover, New York.

GRIEG, D.M. 1980. *Optimisation*. Longman, London.

LIGHTHILL, M.J. (Ed.). 1978. *Newer Uses of Mathematics*. Penguin, London.

LIPSHUTZ, S. 1966. *Finite Mathematics*. McGraw-Hill, New York.

ORE, O. 1963. *Graphs and Their Uses*. Random House, New York.

SETHI, S.P. and THOMPSON, G.L. 1981. *Optimal Control Theory: Applications to Management Science*. Martinus Nijhoff, Boston, MA.

WILSON, R.J. and WATKINS, J.J. 1990. *Graphs, An Introductory Approach*. Wiley, New York.

Answers to Exercises

Chapter 1

1.1 £160.64

1.2 £2493.96

1.3 $a = -1, b = 1$; $[0, -12]$

1.4 £162.89

1.11 $A = \begin{bmatrix} 0.9 & 0.2 \\ 0.1 & 0.8 \end{bmatrix}$

1.13 (a) $7(2)^k - 4$ (b) $7 - 2k$

1.14 $x_k = \left(\frac{n-2}{n-1}\right)x_{k-1} + 2s$

1.16 (a) $\frac{17}{4}(3)^k - \frac{1}{2}k - \frac{1}{4}$

(b) $\frac{11}{9}(-2)^k - \frac{1}{3}k + \frac{7}{9}$

(c) $\frac{1}{2}k^2 + \frac{3}{2}k - 1$

1.17 (a) $5.995\pi k + 0.005\pi k^2$

(b) 778.49 cm (c) 3.055 cm

1.18 (a) $11(-3)^k - 9(-4)^k$

(b) $3(2+i)^k - (2-i)^k$

1.19 $6(8 - 3k)(5)^{k-1}$

1.20 $c_1 \cos k\alpha + c_2 \sin k\alpha$; $a = 1$: $c_1 + c_2 k$; $a = -1$, $(c_1 + c_2 k)(-1)^k$

1.22 (a) $\frac{4}{7}(-9)^k - \frac{18}{7}(-2)^k + 1$

(b) $\frac{19}{4} + \frac{1}{2}k - \frac{3}{4}(-1)^k$

(c) $\left(-1 + \frac{11}{3}k\right)\left(\frac{3}{2}\right)^k + 1$

1.24 (a) $\dfrac{z}{(z-1)^2}$ (b) $\dfrac{z}{z-e^c}$

1.25 (a) $\dfrac{z(2z+17)}{z^2+7z+12}$ (b) $\dfrac{z(2z-4+4i)}{z^2-4z+5}$

(c) $\dfrac{-2z(z+4)}{(z+3)^2}$

1.26 (a) $-\frac{1}{7}\left[(-9)^k-(-2)^k\right]$ (b) $7(2)^k-4$ (c) $7-2k$

1.28 (a) $-\frac{7}{9}(-2)^k-\frac{1}{3}k+\frac{7}{9}$

1.31 (a) $a=\frac{3}{2}, b=\frac{11}{12}, c=\frac{1}{12}$

$x_k=\frac{11}{12}+\frac{1}{12}(-5)^k+\frac{3}{2}k$

1.32 $\frac{1}{2}ck(k-1)$

1.33 $a=5.995\pi, b=0.005\pi$

1.35 (a) $\begin{bmatrix} 6 \\ 11 \\ -16 \end{bmatrix}$ (b) $\begin{bmatrix} -12 \\ 48 \\ -5 \\ 33 \end{bmatrix}$

1.36 $\begin{bmatrix} 2550 \\ 1475 \\ 125 \end{bmatrix}, \begin{bmatrix} 4925 \\ 637.5 \\ 737.5 \end{bmatrix}$

1.37 (a) $\begin{bmatrix} 6 & 21 & 31 \\ 11 & -15 & 45 \\ -16 & 34 & -57 \end{bmatrix}$ (b) $\begin{bmatrix} -12 & -4 & -24 & 19 \\ 48 & 42 & 69 & 65 \\ -5 & 9 & -3 & -6 \\ 33 & 45 & 8 & 58 \end{bmatrix}$

1.39 $-2, 4; \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

1.41 36, 54

1.43 $\begin{bmatrix} i \\ 1 \\ -1 \end{bmatrix}, \begin{bmatrix} i \\ -1 \\ 1 \end{bmatrix}$

1.44 $\begin{bmatrix} 1 \\ \frac{1}{6} \\ \frac{1}{18} \end{bmatrix}; -1, -\frac{1}{2}$

1.46 $L=\begin{bmatrix} 0 & 7 & 6 \\ \frac{1}{16} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \end{bmatrix}$

1.48 $X(10) = \begin{bmatrix} 666\left(\frac{3}{2}\right)^{10} + 216\left(-\frac{1}{2}\right)^{10} - 736 \\ 111\left(\frac{3}{2}\right)^{10} - 108\left(-\frac{1}{2}\right)^{10} + 184 \\ 37\left(\frac{3}{2}\right)^{10} + 108\left(-\frac{1}{2}\right)^{10} - 92 \end{bmatrix}$

1.49 $\frac{1}{2}\begin{bmatrix} (4^{100} + 2^{100}) & (4^{100} - 2^{100}) \\ (4^{100} - 2^{100}) & (4^{100} + 2^{100}) \end{bmatrix}$

1.50 $A^k = \frac{2}{3}\begin{bmatrix} 1 - \left(-\frac{1}{2}\right)^{k+1} & \frac{1}{2} + \left(-\frac{1}{2}\right)^{k+1} \\ 1 - \left(-\frac{1}{2}\right)^{k} & \frac{1}{2} + \left(-\frac{1}{2}\right)^{k} \end{bmatrix}$

1.51 5.27 : 1.99 : 1; 7.02 : 1.77 : 1

A2 $S_n = n^2$

Chapter 2

2.3 error, 00110, error, error

2.4 3

2.5 Errors in even-numbered digits of 5 in first case or 2, 4, 6 or 8 in second case are not detected. All transposition errors detected in first case, but not in second case if $x_i - x_{i+i} = \pm 5$.

2.6 (i) 7 o'clock (ii) 7 o'clock (iii) 8 o'clock

2.8 All transpositions are detected (note: for $x_4 \leftrightarrow x_5$, $x_5 \leftrightarrow x_6$ visual inspection of passport holder detects $|x_4 - x_5| = 5$, $|x_5 - x_6| = 5$).

2.9 (a) 2 (b) 3

2.10 (a) 101010 (b) 010101 (c) more than one error

2.11 $d = 5$, detects four errors

2.13 (a) no (b) yes (c) yes

2.14 (a) $d = 4$ (b) $d = 2$

2.15 (a) $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ (b) $\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$

2.16 0000000, 1111001, 1110010, 1010100, 0100110, 0001011, 0101101, 1011111

2.17 0011010, 0101001, 1010110, 1100101, 0110011, 1001011, 0011101, 0101110, 0000111, 1111111

2.19 (a) 01001 (b) 11110, 10111

2.20 (a) five (b) six

2.21 three check bits

2.22 (a) six (b) 10 (c) many possible choices

2.23 (a) $\begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$ (b) 10110, 11101, 10110

2.24 (a) 011001, 110101

(b) 001111, 110101 (two errors); 101100 (three errors)

2.27 (b) (i) 0010011101 (ii) 1001010011
(c) (i) 001001 (ii) 010101 (iii) more than one error

2.28 11101100010

2.31 (a) no (b) yes

2.32 check digit = 5

2.33 6

2.34 (b) 0 (c) 0471621773, 0481621873, 0471623873

2.38 0206241909, 0612960587

2.42 4003101715

Chapter 3

3.1 £443.69

3.9 $A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & -\mu/m & 0 \\ k\rho & 0 & -k \end{bmatrix}, B = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$

3.10 $A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1.1g & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0.1g & 0 & 0 & 0 \end{bmatrix}, B = \begin{bmatrix} 0 \\ -1 \\ 0 \\ 1 \end{bmatrix}$

3.11 $A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0.12 & 0.95 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0.14 & 0 & 0 & 0.95 \end{bmatrix}, B = \begin{bmatrix} 0 \\ -1 \\ 0 \\ 0 \end{bmatrix}$

3.12 (c) $u(1) = -1$, $u(0) = 3$

3.13 $\alpha = 1$ or 2

3.14 not controllable

3.17 $a = 0$, 1 or 3

3.18 controllable

3.19 (b) $x(0) = -\frac{1}{40}\begin{bmatrix} 73 \\ 91 \end{bmatrix}$

3.20 $\beta = -1$ or $-\frac{1}{2}$

3.21 observable

3.23 yes, observable

3.24 yes, observable

3.25 $x(0) = \begin{bmatrix} 3 \\ -1 \end{bmatrix}$

3.26 $f = \left[\frac{23}{19}, -\frac{33}{19}\right]$

3.27 $f = [-19, -6, -11]$

3.29 second fixed eigenvalue $= -1$; $f_3 = f_1 - 1, f_1$ arbitrary

3.30 $f_1 = -4, f_2 < -\frac{7}{2}$

3.31 (a) $k = 10$ (b) $k \neq 10$ (c) impossible

3.32 (a) 3 (b) 2 (c) 3

3.33 controllable

3.34 controllable when $u_1 = 0$, not controllable when $u_2 = 0$

3.35 $a = 0$ or 1

3.37 63

3.39 observable

Chapter 4

4.1 $x(250 - x)$

4.2 $4x^3 - 300x^2 + 5000x$, $0 \leqslant x \leqslant 25$

4.3 $0.7143 \leqslant x^* \leqslant 0.80955$

4.4 $N = 11$

4.5 $1.882 \leqslant x^* \leqslant 2.058$

4.6 122.8×127.2

4.7 Maximize $z = 50x_1 + 50x_2$
subject to: $x_1 + 2x_2 \leqslant 80$, $3x_1 + 2x_2 \leqslant 120$; $x_1 \geqslant 0, x_2 \geqslant 0$

4.9 $x_1 = 20, x_2 = 30, z_{max} = £2500$

4.10 $x_1 = 40, x_2 = 0, z_{max} = £3200$

4.11 $x_1 = \frac{9}{4}, x_2 = \frac{17}{8}, z_{\max} = \frac{87}{8}$

4.12 $x_1 = \frac{33}{5}, x_2 = \frac{4}{5}, z_{\min} = \frac{107}{5}$

4.13 (a) $x_1 = \frac{7}{3}, x_2 = \frac{2}{3}, z_{\max} = \frac{5}{3}$

(b) $x_1 = \frac{1}{3}, x_2 = \frac{8}{3}, z_{\min} = -\frac{7}{3}$

4.14

5		
10	5	
	10	
	10	5

4.17 $x_{ij} = 1$ if athlete i runs in race j; $= 0$ otherwise.

c_{ij} = times in table

Minimize (4.43) subject to (4.41) and (4.42).

4.18 Subjects: Mathematics, Science, Economics, History, French
Teachers: $T_2\,T_1\,T_3\,T_4\,T_5$; or $T_3\,T_1\,T_4\,T_5\,T_2$

4.19 (a) cycle (b) walk from A_1 to A_4 (c) path from A_5 to A_2

4.21 39

4.22 *s*345*f*

4.23 Shortest paths are *s*345*f* or *s*34*f*.

4.24 (a) 14 (b) 19 (c) 26

4.25 white

4.26 Total of 19 matches: four in round 1, eight in round 2, four in round 3, two in round 4, one final.

4.27 10 different outcomes

4.30 Bradford–Manchester–Sheffield–York–Newcastle; or Bradford–Newcastle, Bradford–Manchester, Bradford–Sheffield, Bradford–York
minimum length = 325 km

4.32 {(1, 2), (2, 7), (6, 7), (7, 5), (5, 4), (4, 3)}
{(2, 1), (1, 6), (6, 7), (7, 5), (5, 4), (4, 3)}
length = 130

4.33 $\begin{bmatrix} 14 & 8 \\ 9 & 4 \end{bmatrix}$

4.35 $A^3 = \begin{bmatrix} 4 & 5 & 5 & 5 \\ 5 & 2 & 5 & 2 \\ 5 & 5 & 4 & 5 \\ 5 & 2 & 5 & 2 \end{bmatrix}$

4.36 $A^2 = \begin{bmatrix} 3 & 1 & 2 & 2 & 1 \\ 1 & 2 & 1 & 1 & 2 \\ 2 & 1 & 4 & 1 & 2 \\ 2 & 1 & 1 & 2 & 1 \\ 1 & 2 & 2 & 1 & 3 \end{bmatrix}$

4.37 $J = \frac{9}{5}$; $J = n^2/(2n-1) \to \infty$ as $n \to \infty$

4.38 $u^* = -4e^{-t}/(1 - e^{-2})$

4.40 12

Answers to Problems

Chapter 1

1.1 $r = 6.20$

1.2 $x_k = \left(\dfrac{n-p-1}{n-1}\right)x_{k-1} + (p+1)s$

1.3 $x_0 + k(k+1)(2k+1)/6$

1.4 $x_{k+1} = 0.62x_k,\ y_{k+1} = 0.87y_k + 0.38x_k$

$x_k = (0.62)^k x_0,\ y_k = [5.56(0.87)^k - 4.56(0.62)^k]y_0$

1.11 $s_k = \frac{1}{8}[(75)^{k+1} - (-5)^{k+1}]$

1.13 $x_k = k^2(k+1)^2/4$

1.15 $p_k = (\alpha^k - \alpha^b)/(1-\alpha^b),\ p \neq q;\ 1 - k/b,\ p = q$

1.16 $x_k = c_1(1/\sqrt{2})^k + c_2(-1/\sqrt{2})^k + \frac{1}{2}c[3 - (-1)^k]$

1.17 $\alpha^k x_0 + \dfrac{\alpha^k - c^k}{\alpha - c};\ \alpha^k x_0 + kc^k$

1.19 $Z(i_k) = \dfrac{z^2 - z(1 + V/i_0R)}{z^2 - 3z + 1}\, i_0$

$i_k = i_0 \cosh wk + \dfrac{1}{\sqrt{5}}(i_0 - 2V/R)\sinh wk$

1.21 $p_{k+1} + bcp_k = a$

$p_k = (-bc)^k + a/(1+bc)$

1.24 $x_k = 50,\ y_k = 160$

1.25 $\begin{bmatrix} x_k \\ y_k \end{bmatrix} = \frac{10}{9}\begin{bmatrix} \frac{2}{5} + \frac{1}{2}\left(\frac{1}{10}\right)^k & \frac{2}{5} - \frac{2}{5}\left(\frac{1}{10}\right)^k \\ \frac{1}{2} - \frac{1}{2}\left(\frac{1}{10}\right)^k & \frac{1}{2} + \frac{2}{5}\left(\frac{1}{10}\right)^k \end{bmatrix}\begin{bmatrix} x_0 \\ y_0 \end{bmatrix}$

1.27 (a) $A = \begin{bmatrix} 2 & \frac{16}{3} & 0 \\ 0 & \frac{2}{3} & 2 \\ \frac{1}{3} & 0 & 0 \end{bmatrix}$; $\lambda_1 = \frac{8}{3}$, $\lambda_2 = 2i/\sqrt{3}$, $\lambda_3 = -2i/\sqrt{3}$

1.28 $F_k = (0.096F_0 + 0.85F_1)(1.063)^k + (0.904F_0 - 0.85F_1)(-0.113)^k$

1.30 $x_k = 2^{k+1}$; 8 months

1.31 $x_k = 2^{k+2} - 3$

1.32 option (ii)

Chapter 2

2.1 (a) 6 (b) 51593-2067 (d) 20782-9960
2.2 (c) Non-adjacent transpositions detected only if x_{11} is involved.
2.3 (a) 93471 (c) detected only if $x_i - x_{i+1} \neq \pm 5$
2.4 101101110111101, 00110011000
2.5 0198548273; 0198538723, 0198532873
2.7 positions 3, 5, 9; magnitudes 4, 2, 7
2.9 (b) only detected if $x_i - x_{i+1} \neq \pm 5$
2.10 9770261307057

Chapter 3

3.2 all controllable
3.3 yes, observable
3.4 yes, observable

3.6 $A = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 1 & -\frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$, $B = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$

3.7 $A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -a & 0 \\ d & 0 & -c & 0 \end{bmatrix}$, $B = \begin{bmatrix} 0 \\ 0 \\ b \\ d \end{bmatrix}$

3.8 $k = 20$ or 36
3.9 controllable
3.11 (a) yes, controllable (b) yes, observable
3.13 $k_1 \neq \frac{1}{2}$

3.15 $x_1 = -1 + 2^{k+1} - 3^{k-1}$, $x_2 = 1 - 2^k + 3^{k-1}$, $x_3 = -2^{k+1} + 2(3)^{k-1}$

Chapter 4

4.1 $10.40 \leqslant x^* \leqslant 10.70$

4.2 $1.9 \leqslant x^* \leqslant 2.1$

4.3 $N = 8,\ 3.99 \leqslant x^* \leqslant 4.16;\ N = 7,\ 4.05 \leqslant x^* \leqslant 4.25$

calculus solution = 4.12; actual can has $x \simeq 3.16$

4.4 Maximize $18x_1 + 35x_2$

subject to: $x_1 + x_2 \leqslant 30$

$x_1 + 2x_2 \leqslant 35$

$x_1 \leqslant 19,\ x_2 \leqslant 17,\ x_1 \geqslant 0,\ x_2 \geqslant 0$

Solution is $x_1 = 19,\ x_2 = 8$.

4.5 Minimize $z = 20\,000(x_1 + x_2)$

subject to: $x_1 + 2x_2 \geqslant 80$

$3x_1 + 2x_2 \geqslant 160$

$5x_1 + 2x_2 \geqslant 200$

$x_1 \geqslant 0,\ x_2 \geqslant 0$

Solution is $x_1 = 40,\ x_2 = 20$.

4.6 Maximize $4.1x_{11} + 14.5x_{21} - 7.7x_{12} + 2.7x_{22}$

$+ 10.4x_{31} - 9.1x_{13} + 1.3x_{23} - 1.4x_{32} - 2.8x_{33}$

subject to: $11x_{11} - 9x_{21} - 9x_{31} \geqslant 0$

$x_{11} - 3x_{21} + x_{31} \geqslant 0$

$3x_{12} - x_{22} - x_{32} \geqslant 0$

$3x_{12} - 2x_{22} + 3x_{32} \geqslant 0$

$x_{11} + x_{12} + x_{13} \leqslant 150\,000$

$x_{21} + x_{22} + x_{23} \leqslant 160\,000$

$x_{31} + x_{32} + x_{33} \leqslant 80\,000$

$x_{ij} \geqslant 0$, all i and j

4.7 minimum cost = 535

two solutions

0	0	0	25
15	0	0	10
0	20	30	0

0	0	0	25
0	15	0	10
15	5	30	0

4.8 $x_1 = 0,\ x_2 = \frac{2}{7},\ x_3 = \frac{11}{28},\ x_4 = 0,\ z_{\min} = \frac{4}{7}$

4.9 Minimize $17x_{11} + 19x_{12} + 21x_{22} + 20x_{23}$
$+ 16x_{31} + 15x_{32} + 14x_{33} + 15x_1 + 17x_2 + 10x_3$

subject to:
$$x_{11} + x_{21} \leqslant 15$$
$$x_{22} + x_{32} \leqslant 21$$
$$x_{13} + x_{23} + x_{33} \leqslant 20$$
$$20x_{11} + 17x_{12} + x_1 = 290$$
$$18x_{22} + 15x_{23} + x_2 = 200$$
$$19x_{31} + 18x_{32} + 17x_{33} + x_3 = 230$$
$$x_{ij} \geqslant 0, x_i \geqslant 0, \text{ all } i \text{ and } j$$

4.10 Minimize (4.43) subject to (4.41) and (4.42), where $x_{ij} = 1$ if horse j has rider i; $= 0$, otherwise; c_{ij} = penalty points.

4.11 Minimum number of extension leads is seven.

4.12 Nine different outcomes; probability of one red, one black $= \frac{4}{17}$.

4.13 Eleven different outcomes; three involve less than five games.

4.15 Minimum length of cable is 758 km.

4.16 $A^2 = \begin{bmatrix} 3 & 1 & 3 & 1 \\ 1 & 2 & 0 & 3 \\ 3 & 0 & 5 & 0 \\ 1 & 3 & 0 & 5 \end{bmatrix}$

4.18 ABED, AAECCDB

Index

Some Modern Applications of Mathematics

Emphasizing discrete models using difference equations and matrix representations, this book plays down the importance of calculus and differential equations. Realising that many students are not attracted to traditional applied mathematics, with its bias towards mechanics, the author uses modern and interesting illustrative examples.

KEY FEATURES

Contains a unique combination of topics, including error-correcting codes, optimization, and control theory.

Focuses on practical applications in business, commerce, information technology and the environment - for example, understanding supermarket bar codes or planning a cable TV network.

Provides numerous worked examples and class-tested problems throughout, complete with answers.

Uses an informal and readable approach, so as to be accessible to a wide range of students.

Written by a well-known authority in the field.

Stephen Barnett has been researching, writing and teaching in the areas covered by this book for over 35 years. He is currently an Honorary Professor in the Department of Applied Mathematical Studies at the University of Leeds. Professor Barnett gained his PhD from Loughborough University and his DSc from the University of Manchester; he is a Chartered Engineer. He has published over 120 research papers and seven books in the area of applied mathematics.

ISBN 0-13-834094-3

9 780138 340940